A review on dimensionality reduction for multi-label classification

Wissam Siblini, Pascale Kuntz, Frank Meyer

A Review on Dimensionality Reduction for Multi-label Classification

Wissam Siblini, Pascale Kuntz and Frank Meyer

Abstract—Multi-label classification has gained in importance in the last decade and it is today confronted to the current needs to process massive raw data from heterogeneous sources. Therefore, dimensionality reduction, which aims at reducing the number of features, labels, or both, knows a renewed interest to enhance the scaling properties of the classifiers and their predictive performances. In this paper we review more than fifty papers presenting dimensionality reduction approaches for multi-label classification and we propose an analysis in three steps : (i) a typology of the methods describing the main components of their strategies, the problem they tackle and the way they solve it (ii) a unified formalization of the problems to help to distinguish the similarities and differences between the approaches, and (iii) a meta-analysis of the published experimental results inspired by the consensus theory to identify the most efficient algorithms.

Index Terms—Dimensionality reduction, multi-label classification, meta-analysis.

1 Introduction

The most popular classification paradigms are the single label classification and the multi-class classification. For the first one, the objective is to decide, for each instance described by its features, whether it is associated to a given label or not. The second one is a generalization and it aims at associating each instance to one label among several. However, in many real-world applications (e.g. sound analysis [1] [2], computer vision [3] [4], text analysis [5] [6], biology and health [7] [8], recommender systems [9] [10]), items are intrinsically describable with multiple labels. For instance, in a Video on Demand catalog, a movie is described by a set of complementary labels (e.g. *Funny*, *Masterpiece*, *Based on novel*, *Futuristic*) which are used by a recommender system to provide users with movies that are relevant to their preferences. Consequently, multi-label classification, which associates each instance to multiple labels, has received a great attention in recent years. From the pioneering works of Boutell and al. [11], Zhang and al. [12] and Tsoumakas and al. [13], several reviews have been published [13] [14] [15] [16] [17] [18]. They group the algorithms in three main families : (i) the problem transformation methods which transform the multi-label problem into one or several single-label classification or regression problems, (ii) the algorithm adaptation methods which adapt existing algorithms to learn from multi-label data and (iii) the ensemble methods which deduce multi-label predictions from a collection of learners.

This effervescence in research has allowed a significant improvement of the result quality for benchmarks routinely used in the literature. But it has also coincided with the explosion of data dimensionality. In particular, today, the expansion of online labeling services generates a production of massive raw data of varying quality. This scaling evolution has recently led to the emergence of the so-called eXtreme Multi-label Learning community which considers problems in which the number of labels is extremely large (in the order of 10^6 and more) [19] [20] [21]. This increasing complexity entails a renewed interest for the dimensionality reduction approaches which aim at reducing the number of features, labels, or both in order to improve the scaling properties of the classifiers and their predictive performances.

Dimensionality reduction has a long history in data science [22] [23] associated to different motivations such as, in particular, data visualization and interpretation [24], data compression [25] and data denoising [26]. In short, applying dimensionality reduction on raw data offers a synthetized representation which allows highlighting links and structures hidden in the mass and guiding learning algorithms [27] [28]. As a promising lever for dealing with large and noisy data, dimensionality reduction in multi-label classification has been the subject of a large number of publications over the last decade, resulting in various developments of methods. However, to the best of our knowledge, only one state-of-the-art was already published five years ago [29] and it neither explores the wide range of existing approaches nor provides a global framework to compare them.

For the study presented in this paper, we have gathered more than fifty papers to provide a macroscopic view of the dimensionality reduction strategies developed in multi-label classification and to help users select the most efficient ones. Let us note that we do not consider the variable selection methods (see [30] for a recent review) which are efficient to change the relative importance of variables but which are not designed to extract semantic links between variables as pointed out by several authors [31] [32]. Here we go beyond a classical state-of-the-art often based on an organized list of the existing works by structuring our analysis of the literature along three complementary objectives : (1) a typology of the different approaches, (2) a unified formalization of the problems, and (3) a meta-analysis of the published experimental results. The typology is built from the main components which determine the nature of the problem and the way to solve it: (i) the choice of the reduced space (feature space, label space or both), (ii) the independence/dependence between the dimensionality reduction

objective and the classification objective, (iii) the characteristics of the transformations which reduce the initial spaces, and (iv) the regularization functions and set of constraints which improve the problem solving process. To help to distinguish the similarities and differences between the approaches with more precision, we introduce two generic formulations which scan the large majority of the problems encountered in the literature. We complete this thorough review of the problem ingredients by a meta-analysis of the experimental comparisons carried out in the papers inspired by the consensus theory [33] [34]. For each selected evaluation measure, the published pairwise comparisons (algorithm A_i is better than algorithm A_j at a statistical significance level α) are represented by a multigraph where the vertices are the algorithms and the directed edges represent the domination relationships extracted from the published experimental results. The analysis of the multigraphs allows to identify communities which are families of algorithms that have been mostly examined separately in the litterature. Moreover, in each community, the approaches which outperform the others are highlighted.

2 Typology of Multi-Label Dimensionality Reduction Methods

Throughout the paper we consider a dataset with N instances described by a set of n_x features and labeled by a set of n_y labels. We denote by X (resp. Y) the $N \times n_x$ (resp. $N \times n_y$) matrix describing the features (resp. the labels). As usually done in the literature, X (resp. Y) also refers to the feature (resp. label) space when there is no ambiguity. The objective of the multi-label classification is to predict the right label vector $y \in \mathbb{R}^{n_y}$ for any feature vector $x \in \mathbb{R}^{n_x}$. During a training phase, given the feature matrix X, a classifier is adjusted to fit its prediction to the label matrix Y.

The vast majority of multi-label classification approaches based on dimensionality reduction follows a two-step process : (1) reduction of X or Y or both, (2) prediction of the labels from the reduced spaces with a classifier. The dimensionality reduction is very often applied as an independent data pre-processing before prediction, but recent research stimulates exploration of the coupling between reduction and classification [35]. Whatever the strategy, the impact of the reduction on the classifier performances is *in fine* evaluated by the quality of the label prediction for which numerous measures have been proposed in the literature (e.g. Hamming Loss, F1) [14] [17]. Consequently, three ingredients are considered in the dimensionality reduction problem: the objective function f_d for the dimensionality reduction which is independent from or dependent on the classifier, the objective function f_c associated to the classifier and the final prediction quality measure m_q. Finally, the choice of the reduced space closely determines the nature of the problem and the way to solve it.

Let us denote the reduction of X (resp. Y) by the $N \times k_x$ (resp. $N \times k_y$) matrix X' (resp. Y') where k_x (resp. k_y) is the dimension of the reduced space X' (resp. Y'). In practice, the values of k_x and k_y are often fixed a priori (100 and 500 are commonly used values [36] [37]) but different classical strategies can be applied to guide their choice in particular when the reduction method performs an eigendecomposition (e.g. k_x and/or k_y are the number of eigenvalues above a fixed threshold, or necessary to preserve a percentage of the total sum of eigenvalues). There are three different ways to tackle the dimensionality reduction problem for the multi-label classification (figure 1): (i) reduce the feature space X into X' and predict the label matrix Y from the reduced feature matrix X', (ii) reduce the label space Y into Y' and predict the reduced label matrix Y' from the feature matrix X, (iii) reduce both the label and the feature spaces into X' and Y' and predict the reduced label matrix Y' from the reduced feature matrix X'.

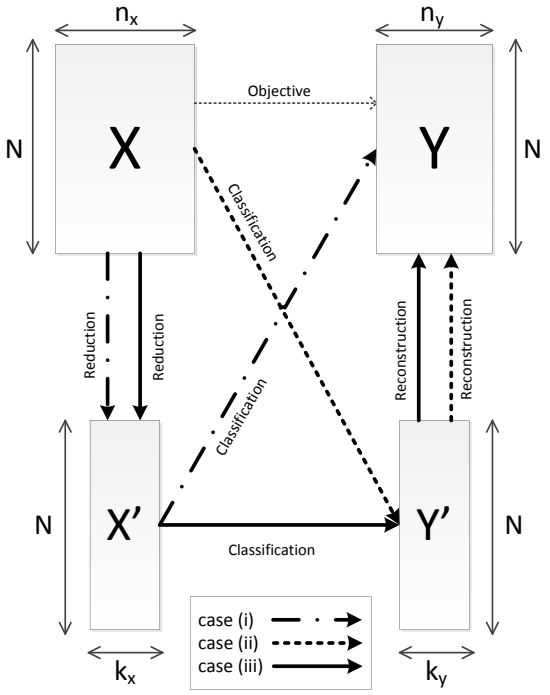

Fig. 1. Overview of the dimensionality reduction strategies in the multi-label classification process.

For each of these cases, the dimensionality reduction problem can be set as an optimization problem:

$$\begin{aligned} \underset{U,V}{\text{optimize}} \quad & f_d(U,V,X,Y) + r(U,V) \\ \text{subject to} \quad & c(U,V,X,Y) \end{aligned} \tag{1}$$

where :

- U and V are either the parameters of a transformation function on X and Y (e.g. projection matrices P_x and P_y in the case of a linear transformation) or the reduced matrices X' and Y'. When a method reduces one space only (either X or Y) the problem is defined with one parameter (U or V).
- f_d is the reduction objective function which is independent from or dependent on the classifier.
- r is a regularization function often associated to a norm (L_1, L_2, L_{12}) on the parameter space which is introduced to limit the overfitting phenomena and to simplify the model.

- c is a constraint set on the search space. Some approaches do not introduce constraints but most of them try to reduce the degree of freedom of the problem to make its resolution easier.

In the following we detail the different ingredients of the problem (1). We first present the most popular objective functions f_d which are independent of the classifiers and we specify their definitions according to the spaces targeted with the dimensionality reduction. Then, we discuss the different cases where the dimensionality reduction objective is coupled with the classification objective. For each case, only one example from the literature is given for illustration and we refer to Table 2 for a detailed state-of-the-art. In addition, Tables 1a and 1b synthetize the strategy of each of the reviewed methods. We finish with a synthetic presentation of the regularization functions and the additional constraints applied by multi-label dimensionality reduction methods.

2.1 Classifier-Independent Objective Functions

We here present the dimensionality reduction methods with an objective function independent of the classifier. They are grouped according to the space they reduce (X, Y, both X and Y).

2.1.1 Feature Space Reduction (X)

The "feature space reduction methods" turn the initial large feature space X into a reduced space X' with the goal of extracting the essential information of the data. As features are partially noisy, redundant and/or irrelevant, some works also aspire to fix the original defects [38]. The objective function f_d is either independent from or dependent on the information carried by the labels.

Most of the label-independent methods have been initially developed for other learning paradigms but quite a few of them have also been frequently applied in multi-label learning. Their objectives can be organized into three families depending on the considered information for the reduction[1]:

1) *Objective FI1*: maximize the conservation of the feature covariance/co-occurrences (e.g. Principal Component Analysis (PCA) [39]);
2) *Objective FI2*: minimize the reconstruction error formulated by a distance between X and X' (e.g. Autoencoders (AE) [40]);
3) *Objective FI3*: maximize the conservation of distances between items described by X and by X' (e.g. Locality Preserving Projection (LPP) [41]). The conservation is either global if all pairwise distances are equally maintained or local if, for example, each item only preserves its distances with its nearest neighbors.

Let us remark that these objectives may be closely linked together; for instance, PCA, classified in FI1, also implicitly minimizes the quadratic reconstruction error between a projection of X' and X (FI2). In addition, besides these approaches, random projections (*Objective R*) have been explored [42] [43].

The label-dependent objectives aim at guiding the reduction with label information [44] [45]. This helps to strengthen the link between the extracted reduced feature space X' and the label space Y. They cover three main strategies:

1) *Objective FD1*: maximize the X-Y link via a standard criterion (covariance, Hilbert-Schmidt Independence) (e.g. Multi-label Dimensionality Reduction via Dependence Maximization (MDDM) [35]) ;
2) *Objective FD2*: preserve the isometry between the instances described in the initial label space Y and the instances described in the reduced feature space X' (e.g. Hypergraph Spectral Learning (HSL) [46])
3) *Objective FD3*: maximize the link between the feature and the label space by learning a subspace X' that can be used to reconstruct both X and Y (e.g. Multi-label Latent Semantic Indexing (MLSI) [47]).

In addition, several hybrid approaches optimize a parameterized trade-off (e.g θ_1 *objective FD1* + θ_2 *objective FD2*) between the above objectives (e.g. Maximizing feature Variance and feature-label Dependence simultaneously (MVMD) [48]).

2.1.2 Label Space Reduction (Y)

As some labels are correlated, it seems intuitive to take these correlations into account to improve both quality and scalability of the classification [17]. This can be achieved by learning a dimensionality-reduced label space. One of the first label space reduction for multi-label classification was based on compressed sensing (CS) [25]. The transformation made by CS is a random projection without training (*Objective R*). However in the prediction phase, CS solves an optimization problem for each instance to reconstruct the label vector from a reduced one. Since then, various strategies have been proposed. Dually to the feature space reduction above, they are either independent from or dependent on the information carried by the features. And, the feature-independent objectives can be organized into three families similar to the label-independent feature space reduction:

1) *Objective LI1*: maximize the conservation of the label covariance (e.g. Principal Label Space Transformation (PLST) [49] which is the equivalent of PCA applied to the label space);
2) *Objective LI2*: minimize the reconstruction error formulated by a distance between Y and Y' (e.g. Multi-label prediction via compressed sensing (CS) [25]).
3) *Objective LI3*: maximize the conservation of distances between items described by Y and by Y' (e.g. Cost-sensitive Label Embedding with Multidimensional Scaling (CLEMS) [50])

Let us note that, similarly to PCA, PLST also implicitly minimizes the quadratic reconstruction error (LI2).

There are also feature-dependent objectives. Indeed, when considering dimensionality reduction for classification, reducing the labels while strengthening the links with

1. Each objective is encoded to be identified in Table 2. For instance, FI1 refers to the **1**st objective of the label-**I**ndependent **F**eature reduction methods.

the features can be useful. Conditional Principal Label Space Transformation (CLPST) [51] is one of the first methods to reduce the labels with an objective dependent on the features and it has opened the way to many other feature-dependent label space reduction approaches. They maximize the correlations between X and Y' to improve the predictability of one matrix from the other (*Objective LD*).

In addition, several hybrid approaches solve a parametrized trade-off between minimizing the reconstruction error between Y and Y' and maximizing the prediction of the feature matrix X from the reduced label matrix Y' (e.g. Dependence Maximization based Label space dimensionality Reduction (DMLR) [52]).

Note that when the label space is reduced and the classification model is trained on Y', the latter predicts reduced label vectors y' and it is necessary to reconstruct the original label vectors y from it. Three cases are commonly encountered in practice (the main associated methods are indicated in parentheses):

1) A reconstruction model $\Psi_{inv} : y' \mapsto y$ is trained during the reduction phase or after [53]. It allows to reconstruct y from y' in the test phase. When the reduction is based on an orthogonal projection P_y, the reconstruction is often computed by the transpose projection ($y' \mapsto y = P_y^T y'$). (PLST, MOPLMS, ML-CSSP, CPLST, BML-CS, FaIE, Rembrandt, DMLR, TRANS, LEML, Bi-Dir, BMLPL, GIMC, C2AE, WSABIE, COMB)
2) If the dimensionality reduction method explicitely provides a reduction function $\Psi : y \mapsto y'$, then, given a reduced label vector y', the original label vector y can be recovered by solving the following structured output learning [54] problem:

$$\min_y l(y', \Psi(y)) \tag{2}$$

where l is a loss function. The optimization is often performed with matching pursuit [55] or basis pursuit [56]. (CS, MLC-BMaD, MSE)
3) The nearest neighbors of y' are computed in the reduced training set and y is deduced from the aggregation of their original label vectors [37]. (CLEMS, SSI, SLEEC)

2.1.3 Both Feature Space and Label Space Reduction

When both spaces are reduced, the reduction of each space depends on the other and two main strategies have been investigated:

- *Objective LFD1*: seek the principal directions in both label space and feature space which maximise the linear correlations with each other. Originally developed with the popular method CCA (Canonical Correlation Analysis) [23] this approach has led to dozens of extensions in multi-label classification (e.g. an extension with a least square resolution LS-CCA [29], an extension with a sketching technique [57], an output-code extension [58]). Moreover, some methods have extended CCA by combining it with other approaches (e.g. The Two-Stage Dual Space Reduction Framework (2SDSR) [59]).
- *Objective LFD2*: minimize a distance function between X' and Y' (e.g. Supervised Semantic Indexing SSI [60]).

In addition, note that there is a special case (Independent Dual Space Reduction (IDSR) [61], [62]) where the label and feature space reductions are independently operated (*objective LFI*): a label-independent feature space reduction is applied on X and a feature-independent label space reduction is applied on Y.

2.2 Coupling Dimensionality Reduction with the Classifier Objective

As previously pointed out, a large majority of the reduction approaches are applied as a data pre-processing independent of the classification stage. But this procedure can turn out to be lacking in flexibility in some cases: its performances may be high for some problems and degrade some others. Indeed, it has been observed on many benchmarks that the impact of a reduction method on the classification performances varies with the classifier and the datasets [48]. To overcome this limitation, some works have started investigating the coupling between dimensionality reduction and classification. At first glance, this approach consists in setting the coupling as a multi-objective optimization problem which tries to optimize both the reduction and the classifier objectives (resp. f_d and f_c) simultaneously. This multi-objective/multi-parameter problem is difficult to solve [63] [64] [65] and, in practice, f_d and f_c are alternatively or jointly maximized via a linear combination (*Objective C1* - e.g. Simultaneous Large-margin and Subspace Learning Approach (TRANS) [66]). But, when we get down to the details, the coupling can also be set up in two other scenarios:

1) *Objective C2*: the dimensionality reduction is integrated within the classification model by replacing X and Y by X' and Y' in f_c and the objective is consequently the maximization of f_c (e.g Linear Dimensionality Reduction for Multi-label Classification (MLSVM) [67]).
2) *Objective C3*: the dimensionality reduction objective f_d is implicitly designed to optimize the classifier. This happens when the classifier is k-NN. For instance, Supervised Orthonormal Locality Preserving Projection (SOLPP) [68] learns a projection P_x on the feature space X that reduces the distance between instances which share numerous labels. This implicitly optimizes k-NN. Similar strategies are employed in other methods (e.g. Hypergraph Spectral Learning (HSL) [46]).

2.3 Explicit and Implicit Transformations

When the algorithm reduces the data via a transformation function, the reduction is explicit and it allows to compute the transformation of any instance on line. Otherwise, the transformation is implicit: it directly provides the reduced matrix but not the transformation function.

#	Method	Description
1	Principal Component Analysis (PCA)	Eigendecomposition of the feature covariance matrix to derive orthogonal directions of maximal variance (principal components).
2	Locality Preserving Projection (LPP)	Spectral decomposition of the instance adjacency graph to compute a reduced feature space that maximally preserves it.
3	Constrained Non-negative Matrix Factorization (CNMF)	Constrained non negative matrix factorization on X. The first factor is considered as the reduced feature space X'.
4	Random Principal Component Analysis (RPCA)	Randomized algebra technique (much faster than PCA when n_x is large) to approximate the principal components of X.
5	Auto-Encoder (AE)	Non linear reduction (projection + activation) and decoding to efficiently reduce and reconstruct the original feature space.
6	Model-Shared Subspace Boosting (MSSBoost)	Multiple random reductions of the feature space to create and combine a set of weak classifiers.
7	Orthonormal Locality Preserving Projection (OLPP)	Extension of LPP with an orthonormality constraint on the reduced feature space.
8	Orthonormal Neighborhood Preserving Projection (ONPP)	Orthonormal feature space projection which preserves each item location with respect to its l nearest neighbors.
9	Shared subspace for multi-Label (ML_{LS})	Embedding resulting from a trade-off between the label space and the feature space reconstructions.
10	Partial Least Square (PLS)	Construction of a dimensionality reducing projection that maximizes the correlations between the projected feature space and the label space.
11	Multi-label Latent Semantic Indexing (MLSI)	Linear feature space projection to optimize both the reconstruction of the original feature space and the correlations between the projected feature space and the label space.
12	Orthonormal Partial Least Square (OPLS)	Extension of PLS with an orthonormality constraint on the reduced feature space.
13	Hypergraph Spectral Learning (HSL)	Spectral decomposition of an hypergraph which links instances with many common labels to obtain a reduced feature space that favors locality between them.
14	Joint Dimensionality Reduction and Multi-label Classification (MLSVM)	Simultaneous learning of a feature space reducing projection and an SVM classifier applied on the obtained reduced space.
15	Multi-Label Dimensionality reduction via Dependence Maximization (MDDM)	Linear projection of the feature space that produces a reduced space with a minimal Hilbert Schmidt Independence with the label space.
16	Semi-Supervised Dimension Reduction for Multi-Label Classification (SSDR-MC)	Construction of a feature space projection which reproduces the neighborhood of the instances in the label space in the projected feature space.
17	Supervised Orthonormal Locality Preserving Projection (SOLPP)	Spectral decomposition of a feature/label adjacency trade-off graph to obtain a reduced feature space where neighbors have common features and labels.
18	Multi-label Linear Discriminant Analysis (MLDA)	Proposition of a definition for multi-label interclass/intraclass variances and computation of the linearly reduced feature space that maximizes their ratio.
19	Direct Multi-label Linear Discriminant Analysis (DMLDA)	Redefinition of MLDA's interclass variance matrix to overcome a limit that MLDA has on the dimensionality of the reduced space.
20	Variable Pairwise Constraint projection for Multi-label Ensemble (VPCME)	Proposition of "must/cannot link" constraints between instances (based on their labels) and computation of the feature space projection that maximally respects them.
21	Hypergraph Orthonormal Partial Least Square (HOPLS)	Trade-off between OPLS and HSL.
22	Shared Subspace Multi-Label Dim reduction via Dependence Maximization (SSMDDM)	Trade-off between ML_{LS} and MDDM.
23	Maximizing feature Variance and feature-label Dependence simultaneously (MVMD)	Trade-off between PCA and MDDM.
24	Multi-label prediction via Compressed Sensing (CS)	Reduction of the label space with a random projection and reconstruction of it with a sparse signal identification technique.
25	Principal Label Space Transformation (PLST)	Dually to PCA, computation of the orthogonal directions of maximum variance (principal components) in the label space.
26	Bayesian Multi-Label Compressed Sensing (BML-CS)	Simultaneous learning, with EM, of probabilistic models for (i) labels reduction, (ii) reduced labels prediction and (iii) labels decoding.
27	Multi-Label Classification via Boolean Matrix Decomposition (MLC-BMaD)	Boolean Matrix Decomposition to construct the binary reduced label space that can optimally reconstruct the original label space.

TABLE 1a
Short description of all the algorithms presented in the review.

#	Method	Description
28	Landmark Selection Method for Multiple Output Prediction (MOPLMS)	Resolution of a strongly regularized label space encoding/decoding problem and selection of the non-zero labels in the solution as the reduced label space.
29	Multi-Label Column Subset Selection Problem (ML-CSSP)	Derivation of label weights from the spectrum of the label covariance matrix and label sampling with the weighted probability to produce the reduced space Y'.
30	Cost-sensitive Label Embedding via Multidimensional Scaling (CLEMS)	Multidimensional scaling of the label space to embed instances according to a chosen instance pairwise cost which reflects the similarity of their label vector.
31	Conditional Principal Label Space Transformation (CPLST)	Combination of PLST and CCA to guide label space reduction with feature information.
32	Multi-label Subspace Ensemble (MSE)	Dimensionality reduction of the label space to improve its linear correlations with the feature space.
33	Feature-aware Implicit label space Encoding (FaIE)	Implicit reduction of the labels to maximize (i) their ability to reconstruct the original labels and (ii) their correlation with the feature space.
34	Response EMBedding via RANDomized Techniques (Rembrandt)	Eigenvalue decomposition of the feature-label covariance matrix with a sketching technique to reduce the label space and improve its link with the feature space.
35	Dependence Maximization based Label space dimension Reduction (DMLR)	Trade-off between PLST and MDDM.
36	Multi-Label Adapatative Random Projection (ML-ARP)	Computation of the feature space projection with an RVNS heuristic that optimizes the performances of the multi-label classifier ML-kNN on the reduced feature space.
37	Independent Dual Space Reduction (IDSR)	Independent application of PCA on the feature space and PLST on the label space.
38	Canonical Correlation Analysis (CCA)	Computation of the principal directions in both label and feature spaces that maximizes their linear correlations with each other.
39	Supervised Semantic Indexing (SSI)	Reduction of the feature space and the label space to increase (resp. decrease) the similarity between relevant (resp. irrelevant) pair x'-y' of feature and label vectors.
40	Least-Square Canonical Correlation Analysis (LS-CCA)	Approximate solution of CCA using an equivalent least square expression and an efficient resolution.
41	Regularized Canonical Correlation Analysis (rCCA)	Regularized version of CCA with an improved behavior when the label-feature covariance matrix is close to singular.
42	Web Scale Annotation By Image Embedding (WSABIE)	Construction of feature and label embeddings such as the instances' features average representation is similar to their labels' representation.
43	Simultaneous Large-margin and Subspace Learning Approach (TRANS)	Combination and simultaneous training of an unsupervised feature reduction method and a large margin multi-label classifier.
44	Deep Canonical Correlation Analysis (DCCA)	Application of CCA on $f_1(X)$ and $f_2(Y)$ where f_1 and f_2 are two deep neural networks. CCA and the networks are trained simultaneously.
45	Supervised Dual Space Reduction (2SDSR)	Family of methods that apply an existing dependent feature space reduction method on X (e.g MDDM) and an existing dependent label space reduction method on Y.
46	Convex Co Embedding (ILA)	Projection of the label and feature spaces which optimize a similarity, for each instance, between the reduced feature vector and reduced label vector.
47	Low rank Empirical risk minimization for Multi-Label Learning (LEML)	Simultaneous training of a classifier and a linear feature space reduction with the low rank Empirical Risk Minimization problem.
48	Bi-Directionnal Representation Learning (Bi-Dir)	Simultaneous predictions of labels from features and features from labels with an intermediary dimensionality reduction based on a bi-directional neural network.
49	Bayesian Multi-label Learning via Positive Labels (BMLPL)	EM-based construction of a subset of reduced labels (called topics) that can (i) reconstruct all the labels (Poisson law) and (ii) be predicted from features (Gamma law).
50	Sparse Local Embedding for Extreme Classification (SLEEC)	Clustering of the instances and construction of local embeddings, on each cluster, of the feature space to obtain the same closest neighborhood as in the label space.
51	Multi-label classification with feature-aware non-linear label space transformation (COMB)	Union of a reduced features space obtained with CCA and a reduced feature space obtained with KCCA.
52	Robust Extreme Multi-label Learning (REML)	Prediction of labels using both a linear model applied on a linearly reduced feature space and a sparse linear model applied on the original feature space.
53	Goal Inductive Matrix Completion (GIMC)	Multi-label matrix completion technique to reduce the instance features and labels.
54	Canonical-Correlated Auto-Encoder (C2AE)	Combination of a label space auto-encoder with CCA to reduce the features and the labels and to decode the predicted reduced labels.

TABLE 1b
Short description of all the algorithms presented in the review.

Family of method	Dependancy	Method	Ref	Year	Constraint/ Regularization	Objective Code	Type of transformation	Scale w.r.t N	Scale w.r.t n_x n_y	Coupl. classif
Feature Space Reduction	Label Independent	PCA	[39]	1901	Orth. Transfo.	FI1	Explicit Lin	Yes	No	No
		LSI	[69]	1990	Uncorr. Space	FI2	Implicit	Yes	No	No
		LPP	[41]	2004	Uncorr. Space	FI3	Explicit Lin	No	No	No
		CNMF	[70]	2006	Uncorr. Space	FI1	Implicit	Yes	Yes	No
		RPCA	[71]	2006	Orth. Transfo.	FI1	Explicit Lin	Yes	Yes	No
		AE	[40]	2006		FI2	Explicit Non Lin	Yes	No	No
		MSSBoost	[72]	2007		R	Explicit Lin	No	No	No
		OLPP	**[73]**	2007	Uncorr. Space	FI3	Explicit Lin	No	No	No
		ONPP	**[73]**	2007	Orth. Transfo.	FI3	Explicit Lin	No	No	No
	Label Dependent	PLS	[74]	1983	Orth. Transfo.	FD1	Explicit Lin	Yes	No	No
		MLSI	[47]	2005	Uncorr. Space	FD3	Explicit Lin	Yes	No	No
		OPLS	[75]	2006	Uncorr. Space	FD1	Explicit Lin	Yes	No	No
		HSL	[46]	2008	Uncorr. Space	FD2/C3	Explicit Lin	No	No	Yes
		ML_{LS}	**[76]**	2008	Orth. Transfo.	FD3	Explicit Lin	Yes	No	No
		MLSVM	**[67]**	2009	Orth. Transfo.	C2	Explicit Lin	Yes	No	Yes
		MDDM	**[35]**	2010	Orth. Transfo.	FD1	Explicit Lin	Yes	No	No
		SSDR-MC	[77]	2010	Uncorr. Space	FD2/C3	Implicit	No	No	Yes
		SOLPP	**[68]**	2010	Orth. Transfo.	FD2/C3	Explicit Lin	No	No	Yes
		MLDA	**[78]**	2010	Orth. Transfo.	FD1	Explicit Lin	Yes	No	No
		DMLDA	**[79]**	2013	Orth. Transfo.	FD1	Explicit Lin	Yes	No	No
		VPCME	[80]	2013	Orth. Transfo.	FD2/C3	Explicit Lin	No	No	Yes
		HOPLS	**[81]**	2014	Uncorr. Space	FD1/C3	Explicit Lin	No	No	Yes
		SSMDDM	**[82]**	2015	Regularization	FD1/FD3	Explicit Lin	Yes	Yes	No
		LM-kNN	[83]	2015		C3	Explicit Lin	Yes	No	Yes
		MVMD	**[48]**	2016	Orth. Transfo.	FI1/FD1	Explicit Lin	Yes	No	No
		ML-ARP	**[84]**	2017		C2	Explicit Lin	No	Yes	Yes
Label Space Reduction	Feature Independent	CS	[25]	2009		R	Explicit Lin	Yes	Yes	No
		PLST	**[49]**	2012	Orth. Transfo.	LI1	Explicit Lin	Yes	No	No
		MLC-BMaD	**[85]**	2012	Binary Space	LI2	Implicit	No	No	No
		MOPLMS	**[86]**	2012	Selection	LI2	Explicit	Yes	Yes	No
		ML-CSSP	**[87]**	2013	Selection	LI1	Explicit	Yes	No	No
		CLEMS	[50]	2016		LI3	Implicit	No	Yes	No
	Feature Dependent	CPLST	**[51]**	2012	Orth. Transfo.	LD	Explicit Lin	Yes	No	No
		MSE	**[88]**	2012		C2	Implicit	No	No	Yes
		BML-CS	[89]	2012		C1	Explicit Non Lin	Yes	No	Yes
		FaIE	**[90]**	2014	Uncorr. Space	LD	Implicit	Yes	No	No
		Rembrandt	**[91]**	2015		LD	Explicit Lin	Yes	Yes	No
		DMLR	[52]	2015	Orth. Transfo.	LD	Explicit Lin	Yes	No	No
Feature Space and Label Space Reduction	Independent	IDSR	[61]	2013	Orth. Transfo.	LFI	Explicit Lin	Yes	No	No
	Dependent	CCA	**[92]**	1936	Uncorr. Space	LFD1	Explicit Lin	Yes	No	No
		SSI	[60]	2009		LFD2	Explicit Lin	Yes	Yes	No
		LS-CCA	**[92]**	2011	Uncorr. Space	LFD1	Explicit Lin	Yes	Yes	No
		rCCA	**[92]**	2011	Uncorr. Space	LFD1	Explicit Lin	Yes	No	No
		WSABIE	[93]	2011		LFD2	Explicit Lin	Yes	Yes	No
		TRANS	[66]	2012	Regularization	C1	Implicite	Yes	Yes	Yes
		DCCA	[94]	2013	Uncorr. Space	LFD1	Explicit Non Lin	Yes	Yes	No
		2SDSR	**[59]**	2013	Several	LFD1	Explicit Lin	Yes	No	No
		ILA	[95]	2014		LFD2	Explicit Lin	Yes	Yes	No
		LEML	**[96]**	2014	Regularization	C2	Explicit	Yes	Yes	Yes
		Bi-Dir	**[53]**	2014		C1	Both	Yes	Yes	Yes
		BMLPL	**[97]**	2015		C1	Implicit	Yes	Yes	Yes
		SLEEC	**[37]**	2015	Regularization	C3	Both	Yes	Yes	Yes
		COMB	[98]	2015	Uncorr. Space	LFD1	Explicit Non Lin	Yes	No	No
		REML	**[99]**	2016	Regularization	C2	Both	Yes	Yes	Yes
		GIMC	**[100]**	2016	Regularization	LFD2	Explicit Non Lin	Yes	Yes	No
		C2AE	[101]	2017	Orth. Transfo.	C1	Explicit Lin	Yes	Yes	Yes

TABLE 2
Dimensionality reduction methods, their typological family and criteria. We report "yes" for scalability if the complexity is strictly under quadratic. The objective codes, which refer to the type of objective considered by the methods, are detailed in Section 2.

2.3.1 Explicit Transformations

The vast majority of the methods presented in that review reduce dimensionality with projections ($X' = XP_x$ or $Y' = YP_y$). They are consequently explicit and linear. These linear transformations can be extended to a non linear transformation with the classical kernel trick and most of the linear methods have a kernel extension (e.g. kPCA [102] for PCA, kCCA [103] for CCA).

Additional non linear explicit approaches have been adapted for the multi-label case. They can be classified into three categories:

1) Locally Linear Embeddings [37] [104]: they produce a non linear transformation, deduced from a piecewise linear transformation, by partitioning the label and/or feature space and computing a specific linear transformation per region.
2) Representation learning with neural networks. The target output depends on the network architecture. For the auto-encoders [40] the output is a reconstruction of the input layer. For the multi-label neural networks [105] [106] [107] the output is a prediction of Y (resp. X) and the input is X (resp. Y). More complex architectures, which combine auto-encoders and multi-layer perceptrons, have been recently investigated [101] [53]. For details, we refer to the complete review [27] on representation learning which includes several methods which have been adapted to multi-label classification [108].
3) Probabilistic process [109] [89] [97]. The transformation from the initial space (X or Y) to the reduced space (X' or Y') is a combination of parameterized probability laws (often Normal, Dirichlet and Gamma distributions). In that case, the construction of the reduced space is achieved by inference.

2.3.2 Implicit Transformations

The implicit transformations directly provide the reduced space without explicitely computing the transformation operator (e.g. using Multi-Dimensional Scaling (MDS) [110] or Matrix Factorization [111]). They consequently have no reason to be linear. Direct learning of the reduced spaces X' or Y' offers more degree of freedom in the optimization problem but it is more frequently confronted to overfitting [112]. Moreover, it is not adapted to incremental processes: when a new item is added, the reduction must be fully relaunched. Nevertheless, the recent rise of extreme multi-label classification [19] [113] stimulates the development of implicit transformations [93] [37] [96] [90] [50]. They are adapted to the label space reduction because the online reduction of a label vector is not required. On the contrary, explicit transformations are more suitable for feature space reduction because the transformation of new feature vectors is necessary in the prediction phase.

2.4 Regularization and Constraints

Adding a regularization function r to the objective function f_d or a set of constraints to the optimization problem (1) aims at (i) reducing the degree of freedom of the problem, (ii) providing simpler transformations of the initial spaces into the reduced ones by restricting the parameters, (iii) improving generalization and limiting overfitting and (iv) building more classification-friendly training sets. These objectives, which are common to many machine learning problems, are integrated in the optimization problem in a variety of ways:

- Sparse transformations. Some methods impose sparsity on the reduced space variables or on the reduction function parameters [37] [96]. Formally, sparsity of a matrix is computed from its L_0-norm, but due to its non-continuity and non differentiality the authors usually resort to the L_1-trick and relax the L_0-norm into a L_1-norm [114]. In practice, this approach limits overfitting, optimizes storage and speeds training and prediction up ;
- Limited search space. A major part of the algorithms impose the minimization of the L_2-norm of the parameters. This benefits solutions with low-value parameters [53] [29].
- Sparse and small parameter sets. This is achieved with an Elastic Net Regularization [115] which is a linear combination of L_1 and L_2 regularizations [37].
- Parameter clipping. This regularization restrains the parameter definition domain to a fixed interval with thresholding techniques [116].
- Dropout regularization. Some neural network based approaches regularize their parameters by using the dropout strategy [117] which selects a different random parameters subset at each training step.

Moreover, constraints are also introduced to limit noise and variable correlations which are enemies of most classifiers [118]. Two usual constraints aim at facilitating the classification task :

- Uncorrelated space. Classification is easier when the correlations in the variable space are limited. Such a constraint can be express in the matrix form $X'^T X' = I$ (or $Y'^T Y' = I$). Let us remark that this constraint leads to a L_2-norm regularization ($|X'|_2 = tr(X'^T X) = tr(I)$).
- Orthonormal projection. This constraint is expressed in the popular linear case by $P^T P = I$.

Some authors have also proposed a trade-off $P^T((1-\mu)X^T X + \mu I)P = I$ between these two constraints [48] [35].

3 Two Generic Problem Formulations

The previous section highlights the great variability of the different ways to adress the issue of dimensionality reduction for multi-label classification. In the literature analysis, where each author resorts to his/her own formulation, this variability is an obstacle to a fine understanding of the similarities and differences between the approaches. To make the comparisons easier, we here propose two generic formulations of the general problem (1). The first one, closely linked to a generic scheme of resolution based on eigendecomposition, allows to express more than half of the problems. The second one is an extension which covers all cases. It is associated to a large variety of optimization processes (e.g. gradient descents, Newton method, Lagrangian techniques).

3.1 The Basic Framework

As we show in Table 3, a large number of problems can be written as follows :

$$\begin{aligned} \underset{U}{\text{optimize}} \quad & tr(U^T A_{XY} U) \\ \text{subject to} \quad & U^T B_{XY} U = I \end{aligned} \tag{3}$$

where :

1) A_{XY} and B_{XY} are matrices which are function of X and Y.
2) according to the space that is reduced and the type of transformation, the parameter U is one of the following matrices : X', Y', P_x, or P_y.
3) the optimization goal is either a minimization or a maximization objective.

Problems expressed by (3) can be solved with an eigendecomposition. It is well-known that, using the Lagrangian method [120], the problem (3) is equivalent to optimize λ in the following generalized eigenvalue problem:

$$A_{XY} u = \lambda B_{XY} u \tag{4}$$

The solution U of (3) in the maximization (resp. minimization) case is therefore the matrix of the eigenvectors associated to the k largest (resp. smallest) eigenvalues of (4). In the frequent case where the matrix A_{XY} is symmetric positive, the eigenvectors of A_{XY} can also be retrieved by a singular value decomposition [121] of the square root R_{XY} of A_{XY} defined by $R_{XY}^T R_{XY} = A_{XY}$.

Despite its elegant solution, the eigenvalue decomposition (4) is computationally complex: in the order of n^2 real-valued numbers for spatial complexity and n^3 operations for temporal complexity [122]. For scaling, different approaches are used: fast eigendecomposition techniques (e.g. Jacobi [123] and QR [124]), approximation of the largest eigenvalues (power iteration algorithm [125], Lanczos method [126]), matrix sketching [127] (e.g. in randomized PCA [71] or Rembrandt [91]). In addition, a reformulation of problem (3) in a least square form is also popular to resort to a numerical optimization method (e.g. least square version of CCA [92] or LDA [128]). Indeed, the initial least square form $\min_U \min_M \|R_{XY} - MU^T\|_F^2$, where R_{XY} is the square root of A_{XY}, is equivalent to $\max_U \text{tr}(U^T A_{XY} U)$ with the constraint $U^T U = I$.

For illustration let us consider the classical formulation of PCA $\max_{P_x} \text{tr}(P_x^T (X^T X) P_x)$. Subject to $P_x^T P_x = I$, it can be reformulated into the mean squared reconstruction error minimization problem $\min_{X', P_x} \|X - X' P_x^T\|_F^2$ with simple algebra. The strong constraint $U^T U = I$ is sometimes replaced with a simpler L_2-regularization on U.

Let us note that a portion of the methods expressed with the basic framework (3) are based on graph spectral decompositions [129] [130]. They follow a two-step procedure: (i) build a graph which links the instances with a proximity property (e.g distance on the label space) and (ii) embed the instances in a reduced space by preserving the graph neighborhood structure. The transformation is computed by an eigendecomposition of the normalized Laplacian of the graph (A_{XY} is the normalized Laplacian and B_{XY} the identity matrix).

3.2 Towards a General Framework

The equivalence between the basic framework (3) and a least square formulation highlights both its flexibility and its limits. L_1-regularizations [131], multi-label loss functions other than mean square error [15] and many other items cannot be expressed as a matrix trace. An attempt at generalization has been proposed in [96]. The problem is set as a problem of minimization of the empiral risk (ERM) [132] which does not require a specific loss function nor any specified regularization. Let us denote by $h(x; Z) : x \mapsto \widehat{y}$ the classification model of parameter Z, by $l(y, \widehat{y}) = l(y, h(x; Z))$ the loss function between the predicted label vector $\widehat{y}$ and the true label vector y, and by $r(Z)$ the parameter regularization. The low rank empirical risk minimization problem is expressed as follow:

$$\begin{aligned} \widehat{Z} = \underset{Z}{\text{argmin}} \quad & \sum_{i=1}^{N} \sum_{j=1}^{n_y} l(Y_{ij}, h^j(x_i; Z)) + \lambda r(Z) \\ \text{subject to} \quad & rank(Z) \leq k \end{aligned} \tag{5}$$

Let us remark that this formulation differs from the classical ERM problem: the added rank constraint on $Z \in \mathbb{R}^{n_x \times n_y}$ entails a dimensionality reduction [133].

The formulation (5) covers a large part of the methods of the literature but to include the remaining uncovered cases, we propose a generic formulation of the objective function which is an additive combination of the essential ingredients encountered in the multi-label dimensionality reduction typology:

$$\begin{aligned} J(X', Y', Z_x, Z_y, Z_{xy}) = & \alpha_x e_x(X', X, Z_x) \\ & + \alpha_y e_y(Y', Y, Z_y) \\ & + \alpha_{xy} e_{xy}(X', X, Y, Y', Z_{xy}) \\ & + \alpha_p p(X', Y') \\ & + \alpha_r r(X', Y', Z_x, Z_y, Z_{xy}) \end{aligned} \tag{6}$$

where:

- e_x is a reconstruction error between X and its reduced version X'.
- e_y is a reconstruction error between Y and its reduced version Y'.
- e_{xy} is a joint error between X, Y, X', and Y' which can, for instance, express the classification error.
- r is a parameter regularization.
- Z_x,Z_y,Z_{xy} are the parameters of the reduction and the classification functions.
- p are additional properties imposed on both reduced spaces.

The reconstruction error e_x can be expressed with both the encoding loss l_{x1} (reconstruction of X' from X) and the decoding loss l_{x2} (reconstruction of X from X'):

$$\begin{aligned} \alpha_x e_x(X', X, Z_x) = & \alpha_{x1} l_{x1}(X', f_{x1}(X, Z_x)) \\ & + \alpha_{x2} l_{x2}(X, f_{x2}(X', Z_x)) \end{aligned} \tag{7}$$

where the f functions are parametric models. This is also valid for e_y and e_{xy}.

Method	U	A_{XY}	Dependancy	B_{XY}	Constraint	Year	Ref
PCA	P_x	X^TX	No	I	Orth. Transfo.	1901	[39]
CCAx	P_x	$X^TY(Y^TY)^{-1}Y^TX$	Yes	X^TX	Uncorr. Space	1936	[23] [92]
PLS	P_x	X^TYY^TX	Yes	I	Orth. Transfo.	1983	[74]
KPCA	P_x	$\phi(X)^T\phi(X)$	No	I	Orth. Transfo.	1997	[102]
LPP	P_x	X^TLX	No	XDX^T	Uncorr. Space	2004	[41]
MLSI	P_x	$X^T((1-\theta)X^TX+\theta Y^TY)X$	Yes	X^TX	Uncorr. Space	2005	[47]
OPLS	P_x	X^TYY^TX	Yes	X^TX	Uncorr. Space	2006	[75]
RPCA	P_x	X^TX	No	I	Orth. Transfo.	2006	[71]
KDA	P_x	$S_w^-1S_b$	Yes	I	Orth. Transfo.	2007	[119]
OLPP	P_x	X^TLX	No	I	Orth. Transfo.	2007	[73]
ONPP	P_x	$X^T(I-W)(I-W^T)X$	No	I	Orth. Transfo.	2007	[73]
HSL	P_x	XL_nX^T	No	X^TX	Uncorr. Space	2008	[46]
MLLS	P_x	$S_2^{-1}S_1$	Yes	I	Orth. Transfo.	2008	[76]
MLSVM	P_x	$(X^TX)^\dagger X^TYY^TX$	Yes	I	Orth. Transfo.	2009	[67]
MDDM	P_x	X^THYY^THX	Yes	I	Orth. Transfo.	2010	[35]
MLDA	P_x	$S_w^-1S_b$	Yes	I	Orth. Transfo.	2010	[78]
SOLPP	P_x	$X^T(I-W)(I-W^T)X$	Yes	I	Orth. Transfo.	2010	[68]
SSDR-MC	X'	$X^T(I-W)(I-W)^TX$	Yes	X^TX	Uncorr. Space	2010	[77]
rCCAx	P_x	$X^TY(Y^TY)^\dagger Y^TX$	Yes	X^TX	Uncorr. Space	2011	[92]
CPLST	P_y	$Y^TH(XX^T)^\dagger HY$	Yes	I	Orth. Transfo.	2012	[51]
PLST	P_y	Y^TY	No	I	Orth. Transfo.	2012	[49]
DMLDA	P_x	$X^THYW^{-1}Y^THX$	Yes	I	Orth. Transfo.	2013	[79]
IDSRx	P_x	X^TX	No	I	Orth. Transfo.	2013	[61]
IDSRy	P_y	Y^TY	No	I	Orth. Transfo.	2013	[61]
VPCME	P_x	$S_C-\theta S_M$	Yes	I	Orth. Transfo.	2013	[80]
FaIE	Y'	$YY^T+\theta X(X^TX)^{-1}X^T$	Yes	I	Orth. Transfo.	2014	[90]
FaIE Linear	P_y	$Y^T(Y^TY+\theta X(X^TX)^{-1}X^T)Y$	Yes	Y^TY	Uncorr. Space	2014	[90]
HOPLS	P_x	$X^T(YY^T+\theta S)X$	Yes	X^TX	Uncorr. Space	2014	[81]
DMLR	P_y	$Y^T(I+\theta HXX^TH)Y$	Yes	I	Orth. Transfo.	2015	[52]
MVMD	P_x	$(1-\theta)X^TX+\theta X^THYY^THX$	Yes	I	Orth. Transfo.	2016	[48]

Notations:
$M^\dagger$: pseudo-inverse of a matrix M
L (resp. L_n): graph (resp. normalized) Laplacian
ϕ: kernel transformation
W, S_C, S_M,S_b, S_w, S: pairwise weight matrices
θ, α , β: trade-off parameters
$H=(\delta_{ij}-\frac{1}{N})_{ij}$ where δ is the Kronecker delta
$S_1=I-\alpha T^{-1}$ and $S_2=T^{-1}X^TYY^TXT^{-1}$ where $T=\frac{1}{N}X^TX+(\alpha+\beta)I$

TABLE 3
Connection between dimensionality reduction methods and the basic objective framework (3)

In most cases, the regularization r can be additively decomposed:

$$\begin{aligned}\alpha_r r(X',Y',Z_x,Z_y,Z_{xy}) = \alpha_{r_1} r_1(X') + \alpha_{r_2} r_2(Y') \\ + \alpha_{r_3} r_3(Z_x) + \alpha_{r_4} r_4(Z_y) \\ Y + \alpha_{r_5} r_5(Z_{xy})\end{aligned} \quad (8)$$

In (6), (7) and (8), the α constants are weights that allow trade-offs between the different components of the problem.

All forms of (6) are tackled with customized numerical optimization methods [134] [135]. Considering the convexity, the smoothness, the order, the differentiability and the conditioning of the formulation, the problem is sometimes reformulated (convex relaxation [136], primal/dual conversion [137], preconditioning [138]) and the resolution is either performed with an adapted variant of the gradient descent [139] [140], a coordinate descent [141] or higher order algorithms such as Newton method [142] or Frank Wolfe's algorithm [143]. Also, constrained problems are generally solved with a Lagrangian method [144], with one of its diverse extensions (e.g Augmented Lagrangian like ADMM [145] [37]) or with a projected gradient descent. The choice of the couple formulation/resolution is essential: it affects the spatial and temporal complexities of the computations and the quality of the convergence towards the solution.

To be complete, let us point out that two families of dimensionality reduction methods explored for multi-label classification reach the limits of the generic formulation. The first one includes approaches based on mixture models [97] and solved with the suitable state-of-the-art EM algorithm and its variants [146] [147]. The second one includes the ensemble strategies (bagging [80] [88] and boosting strategy [72]) where multiple dimensionality reducing transformations are trained on bootstraps and aggregated according to two main strategies. Each transformation produces its own reduced space and either the reduced spaces are aggregated into a global reduced space and the classifier is trained on it or a classifier is trained on each reduced space and the predictions of each classifier are aggregated.

4 META-ANALYSIS

Our previous generic frameworks allow to explicitly identify the different ingredients involved in the various ap-

proaches proposed in the literature and to help to understand their common points and differences. However in practice, a question persists: which are the most efficient approaches? It is difficult to answer because only partial comparisons are generally reported in the articles and to the best of our knowledge there exist no experimental studies which compare all the approaches presented in Table 2. M
the computational implementations are very dive
some approaches the source codes or the parame
are even not available. Hence, a normalized c
would entail a recoding of all the algorithms
battery of tests on a unified framework that rer
defined in the research community. With the larg
gorithms to be considered, this would require a c
time-consuming effort, and consequently the e
of the outcomes of the existing published resea
appears as a more realistic alternative. The ex
protocols (datasets, classifiers, performance meas
varying from one publication to another, we here
new meta-analysis methodology.

Often defined as "the statistical analysis of a
lection of analysis results from individual stud
purpose of integrating the findings" [148], the me
has known an increasing development from its
works in the 30's [149] [150]. One of its favor
medicine where aggregation of the available piec
mation is required to make as rational as possible
In computer science, this approach is still unu
great majority of researchers prefer to compare
approaches with a restricted subset of existing
set of benchmarks they habitually use but the fir
(e.g. [151] [152] [153]) seem promising.

In this paper we aim at identifying the dim
reduction approaches used in multi-label classi
which several pieces of evidence show their c
over others: these approaches statistically obta
performances in the results published in the in
conferences and journals with a review process. A
known in multi-label classification that the pe
can be evaluated with a wide range of measures
the relevant methods for each of the most frequ
and independent quality measures. In the follow
present a descriptive analysis of the observed c
and co-occurrences of the algorithms and of the
then we detail the process to extract the dor
proaches, and finally we discuss the obtained res

4.1 Methodology

From all the papers referenced in Table 2, we have
corpus $\mathcal{C}$ of 27 papers –marked in bold type- wh
relevant and exploitable results for a meta-ana
precisely, we have first extracted the 32 papers th... comp...
at least two methods, and then we have removed the 5
papers whose results are given on graphics only since they
are difficult to exploit.

4.1.1 The Considered Algorithm Set

Let us denote by $\mathcal{A}$ the set of the 42 algorithms that appear in the selected papers of the corpus $\mathcal{C}$. The published pairwise comparisons can be described by a multigraph G_c: the vertices represent the algorithms of $\mathcal{A}$ and an edge is added between two algorithms when they are compared in a paper. In the graph layout (Figure 2), the vertex diameter is proportional to the frequency of the associated algorithm and the edge set between a vertex pair is represented by a single edge whose width is proportional to the set cardinality.

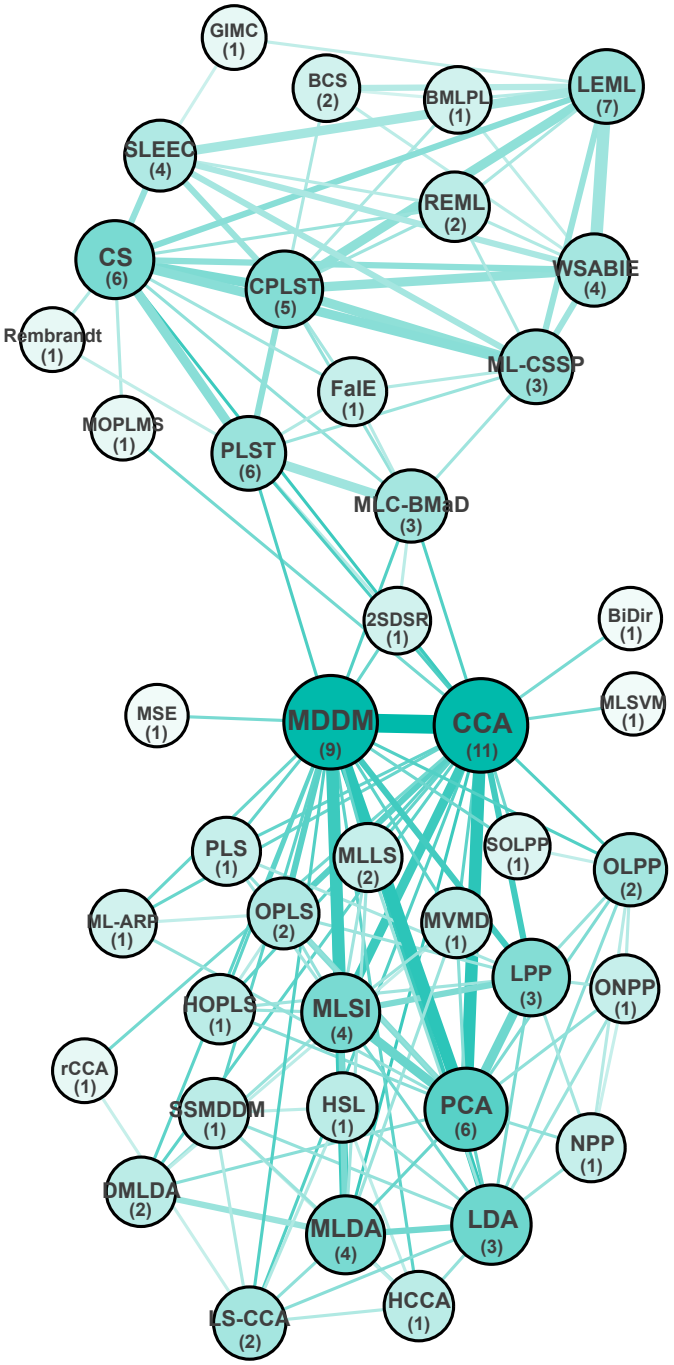

Fig. 2. Algorithm comparison multigraph for the 27 selected articles. The larger and the darker blue, the higher the weights of the edges and degrees of the vertices. The number of articles in which each algorithm appears is in parentheses.

The obtained layout is very different from the one of a complete graph which would be the ideal model, but far from reality, where each algorithm is compared to all others in many experiments. However, it highlights two

	01Loss	AUC	Accuracy	AveragePrecision	Coverage	ErrorRate	F1	HammingLoss	MacroF1	MicroF1	OneError	P@3	Precision	RankingLoss	Recall	SubsetAccuracy	MacroPrecision	MicroPrecision
01Loss	1																	
AUC	0	9																
Accuracy	1	0	3															
AveragePrecision	0	1	0	1														
Coverage	0	1	0	1	1													
ErrorRate	0	0	0	0	0	1												
F1	1	0	3	0	0	0	7											
HammingLoss	1	3	2	1	1	0	3	12										
MacroF1	0	3	0	1	1	0	0	5	7									
MicroF1	0	3	0	1	1	0	0	5	7	7								
OneError	0	3	0	1	1	0	0	2	2	2	5							
P@3	0	2	0	0	0	0	0	1	0	0	4	4						
Precision	0	0	1	0	0	0	2	1	0	0	0	0	2					
RankingLoss	0	1	0	1	1	0	0	1	1	1	1	0	0	1				
Recall	0	0	1	0	0	0	2	1	0	0	0	0	2	0	2			
SubsetAccuracy	0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	1		
MacroPrecision	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	
MicroPrecision	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1

TABLE 4
Matrix of the quality measure co-occurrences computed on the 27-article set. The occurrence of each measure is on the diagonal.

communities that correspond to two families of algorithms which have been mostly studied separately. This confirms that the published comparisons have been done on subsets of algorithms which share common properties. The first community C_1 regroups approaches that reduce the feature space dimension and the second one C_2 regroups label space and co-label and feature space reduction algorithms including those developed in the context of extreme multi-label learning. Moreover, two vertices (CCA and MDDM) appear at the intersection of C_1 and C_2: they have been considered as baselines for a long time and CCA, which reduces both feature and label spaces, naturally belongs to both communities. Three algorithms (BiDir, MLSVM, MSE) are linked to CCA or MDDM only and consequently, in addition to C_1 and C_2, we consider an "in-between subset" C_{1-2} which includes those five algorithms. Edges between C_1, C_2 and C_{1-2} are mainly originated from the reference [62] which is a recent comparison of different multi-label classification approaches. The vertex diameters allow highlighting the most frequently occurring methods which are often mentioned among the pioneers in their community: CCA, MDDM, LEML, PCA, CS, PLST and CPLST. In the following we aim at identifying the significant relationships from the multigraph G_c.

4.1.2 The Evaluation Measures

Table 4 shows the occurrences and co-occurrences of the different measures used in the articles of the corpus $\mathcal{C}$. It underlines the great variability of the considered criteria, and the frequency distribution allows to distinguish the most popular measures: Hamming Loss (44%), AUC (33%), F1 (26%), Macro-F1 (26%), and Micro-F1 (26%). In addition to these observations, our selection of the suitable measures for the meta-analysis is guided by a recent comparison [154] which has experimentally proved that some measures are highly correlated whereas some others are independent. More precisely, their authors have tested a set of 16 measures (those present in Table 4 plus some variants) and have compared them with the Pearson and Spearman correlations on 100 000 simulations. Results show that Hamming Loss, Coverage and Ranking Loss are independent, but here only Hamming Loss is taken into account because the frequency of the two others is very low on $\mathcal{C}$. Results also detect a strong correlation between the measures of a large set $\mathcal{M}$ = {Subset Accuracy or 01Loss, Accuracy, Precision, Recall, F1, One Error, Average Precision, Micro Precision, Macro Precision, Micro F1, Macro F1, Micro Recall, Macro Recall}. Consequently, when several measures of $\mathcal{M}$ are used for a comparison of two algorithms in a same paper, we only retain the most frequent one. Let us precise that AUC and P@3 have not been considered in [154]. But they have been here added to $\mathcal{M}$ as the computation on our data of their Pearson correlation coefficients with the other measures of $\mathcal{M}$ confirms the correlation: its value is ranging between 0.829 (with Macro-F1) and 0.576 (with One-error) for AUC and is close to 1 (with One-error) for P@3. The two studies with Hamming Loss and the subset of selected measures from $\mathcal{M}$ (respectively referred to in the following as $\mathcal{H}$ and $\mathcal{M}$) are conducted separately on the article subsets of $\mathcal{C}$ which take them into account (12 articles for $\mathcal{H}$ and 24 for $\mathcal{M}$).

4.1.3 The Consensus Based Approach

Our meta-analysis inspired by the consensus theory [33] [34] is decomposed into two successive steps: (i) filtering the statistically significant domination relationships for the measures $\mathcal{H}$ and $\mathcal{M}$, and (ii) extracting the dominant algorithms for each measure. Finally we identify the algorithms which statistically dominate the others in the two cases. With a similar process we complete the analysis by distinguishing the algorithms which are dominated.

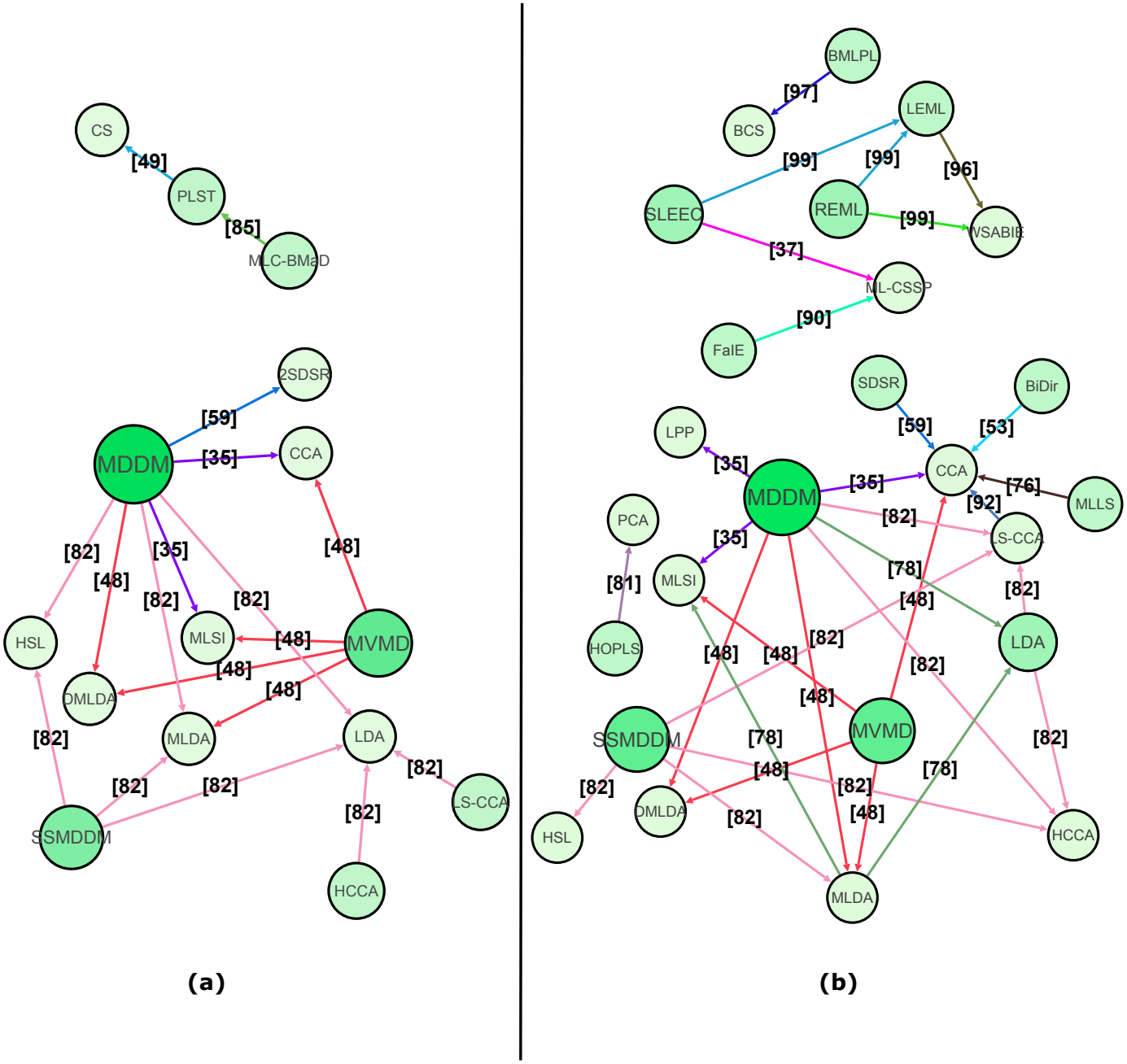

Fig. 3. Domination multigraphs for Hamming Loss (a) and for the set of correlated measures $\mathcal{M}$ (b). Each directed edge is labeled with the reference of the article from which it is extracted. Different colors are associated to different articles.

More precisely, the significant domination relationships are extracted with a Friedman and post-hoc Nemeyi tests [155] with a standard confidence level $\alpha = 0.05$. And, for each measure, we build a directed domination multigraph - denoted respectively by $G_D(\mathcal{H})$ and $G_D(\mathcal{M})$- from G_c by retaining the significant edges and by orienting them according to the direction of the domination: a directed edge from A_i to A_j means that the algorithm A_i significantly outperforms the algorithm A_j in a paper of the corpus $\mathcal{C}$. The first stage of a topological sorting [156] on each multigraph allows identifying the subsets $\mathcal{D}(\mathcal{H})$ and $\mathcal{D}(\mathcal{M})$ of $\mathcal{A}$ which contains the dominant algorithms: A_i is dominant when its indegree is null and its outdegree is strictly positive. Similarly, the dominated algorithms are those with a null outdegree and a strictly positive indegree.

4.2 Results

A multigraph overview reveals communities of algorithms with similar behaviors. A detailed analysis of these communities helps identifying the dominant algorithms.

4.2.1 Algorithm community detection

The two directed multigraphs $G_D(\mathcal{H})$ and $G_D(\mathcal{M})$ are represented in Figure 3. They both have a lot less relationships than the co-occurrence graph G_c. With greater alpha thresholds additional directed edges appear but the confidence that can be placed in them is weaker. As a consequence, for the standard threshold $\alpha = 0.05$, the directed multigraphs

become digraphs with at most one directed edge between each vertex pairs and some algorithms of $\mathcal{A}$ with a null degree are no more represented. Indeed in some articles the number of experiments is too low to detect a domination which is statistically significant. Table 5 indicates, for each article of $\mathcal{C}$ and regardless of the considered quality measure, the ratio CD/$r_{\max}$ between the critical difference CD of the post-hoc Neymeni test for $\alpha = 0.05$ and the theoretical maximal ranking difference $r_{\max}$ between the compared algorithms. If q algorithms are compared, then $r_{\max} = q - 1$. The higher this ratio is, the fewer the expected significant relationships are, and when it is greater than 1 none of them can be extracted.

Ref.	CD/$r_{\max}$	Ref.	CD/$r_{\max}$	Ref.	CD/$r_{\max}$
[35]	0.613	[79]	0.800	[92]	0.576
[37]	0.754	[81]	0.750	[96] 1	0.462
[48]	0.530	[91]	1.657	[96] 2	1.132
[88]	0.653	[49]	0.800	[97]	0.764
[53]	0.877	[51]	0.693	[99] 1	0.867
[67]	1.386	[82]	0.453	[99] 2	0.676
[68]	1.106	[85]	0.800	[100]	0.828
[73]	0.871	[86]	1.172	[87]	0.778
[76]	0.591	[90]	0.750	[84]	0.762
[78]	0.623	[59]	0.672		

TABLE 5
Ratio between the critical difference CD of the post-hoc Neymeni test for $\alpha = 0.05$ and the theoretical maximal ranking difference $r_{\max}$ between the algorithms compared in a given paper. Repeated references are associated to multiple sets of experimental comparisons.

In the multigraphs $G_D(\mathcal{H})$ and $G_D(\mathcal{M})$, the communities identified in sub-section 4.1.1 are associated to different connected components: algorithms from different communities have been infrequently compared and are not linked by a significant relationship. Hence, we present the dominant methods for each community. Results are summarized in Table 6. Due to the bibliographic effect which favors the presence of the best approaches at each period, the most recent (resp. oldest) approaches are more likely to be dominant (resp. dominated) but there are noticeable exceptions such as MDDM and MLLS.

$\mathcal{D}(\mathcal{M})$			$\mathcal{D}(\mathcal{H})$		
C_1	C_{1-2}	C_2	C_1	C_{1-2}	C_2
MVMD	BMLPL	**MDDM**	**MVMD**	MLC-BMaD	**MDDM**
SSMDDM	REML	MLLS	**SSMDDM**		
HOPLS	SLEEC	BiDir	LS-CCA		
LDA	FaIE		HCCA		
	SDSR				

TABLE 6
Dominant algorithms detected with the statistical test with $\alpha = 0.05$ for each measure $\mathcal{H}$ and $\mathcal{M}$ and for each community of the multigraph G_c. Dominant methods for the two measures appear in bold. Different significant thresholds ($\alpha = 0.01$ to 0.1) have been tested and, except for REML for $\alpha = 0.01$, the highlighted algorithms remain at the top.

4.2.2 In-depth comparison

The three methods (MVMD, SSMDDM and MDDM) that dominate for both measures belong to community C_1 (label-dependent feature space reduction methods) or to C_{1-2} and they have close strategies. Let us recall that MDDM minimizes the Hilbert Schmidt Independence Criteria between the reduced feature space and the label space and that MVMD and SSMDDM are hybrid methods whose objective is a trade-off between the objective of MDDM and that of another method (see Table 1a). MVDM and SSMDDM are recent approaches which have been extensively compared to others but in a single paper whereas MDDM, which is older, resists to a larger number of comparisons.

However, to the best of our knowledge, these three methods have not been directly compared to one another. Consequently, in an attempt to better understand their behaviors, we have compared them within the same framework. More precisely, the algorithms have been re-implemented (Python language) and tested in the same computational environment (standard computer with 16Gb of RAM). We used ten multi-label datasets often selected in previous multi-label learning studies. They are divided into train/test sets in the Python library scikit-multilearn[2]. The feature (resp. label) dimensionality varies from 72 to 1836 (resp. from 6 to 983). The reduction methods are combined with the ML-kNN classifier and the parameter settings are extracted from the publications. For each method, ten dimensions k_x have been tested (from 10 to 100 percent of the feature dimensionality n_x) and the best results with the F1-score in the measure set $\mathcal{M}$ and with the Hamming Loss ($\mathcal{H}$) are presented in Figure 4. With the average rank criterion, SSMDMM outperforms MVDM and MDDM for both measures and MVDM is better (resp. worse) than MDDM for $\mathcal{M}$ (resp. $\mathcal{H}$). However, when getting into details, we observe that the results may depend on the datasets: e.g. SSMDMM obtains poor results on the Birds' dataset. This confirms the interest of the meta-analysis which aggregates results gathered on different experiments involving a variety of datasets. Even if a method outperforms the others on average on a limited number of datasets, it is also worth considering those that do not dominate it in a meta-analysis.

The community C_2 is very small in $G_D(\mathcal{H})$: the authors of the algorithms of C_2 are mostly interested in data with a large number of labels and give more importance to ranking measures than to global classification errors such as Hamming Loss. Consequently, there are no methods of C_2 simultaneously dominant for the two measures. For the $\mathcal{M}$ measure, SLEEC and REML dominate in several papers. These methods have been especially designed for extreme multi-label learning, and contrary to the others which build low-rank data representations that can miss the information brought by the long tail label distribution, they compute high-rank representation which capture more useful information. Due to their efficiency, they have gained popularity in recent years.

In addition to these most visible results, it is interesting to identify the approaches that are dominant for one measure and dominated for the other. LS-CCA and HCCA (resp. LDA) are dominant in $G_D(\mathcal{H})$ (resp. $G_D(\mathcal{M})$) and dominated in $G_D(\mathcal{M})$ (resp. $G_D(\mathcal{H})$). These methods are not intrinsically expected to be more efficient for Hamming Loss than for the $\mathcal{M}$ measures, and due to the absence of correlation between $\mathcal{M}$ and $\mathcal{H}$, it is not surprising to find different behaviors. This result confirms the interest of this double analysis.

2. http://scikit.ml/

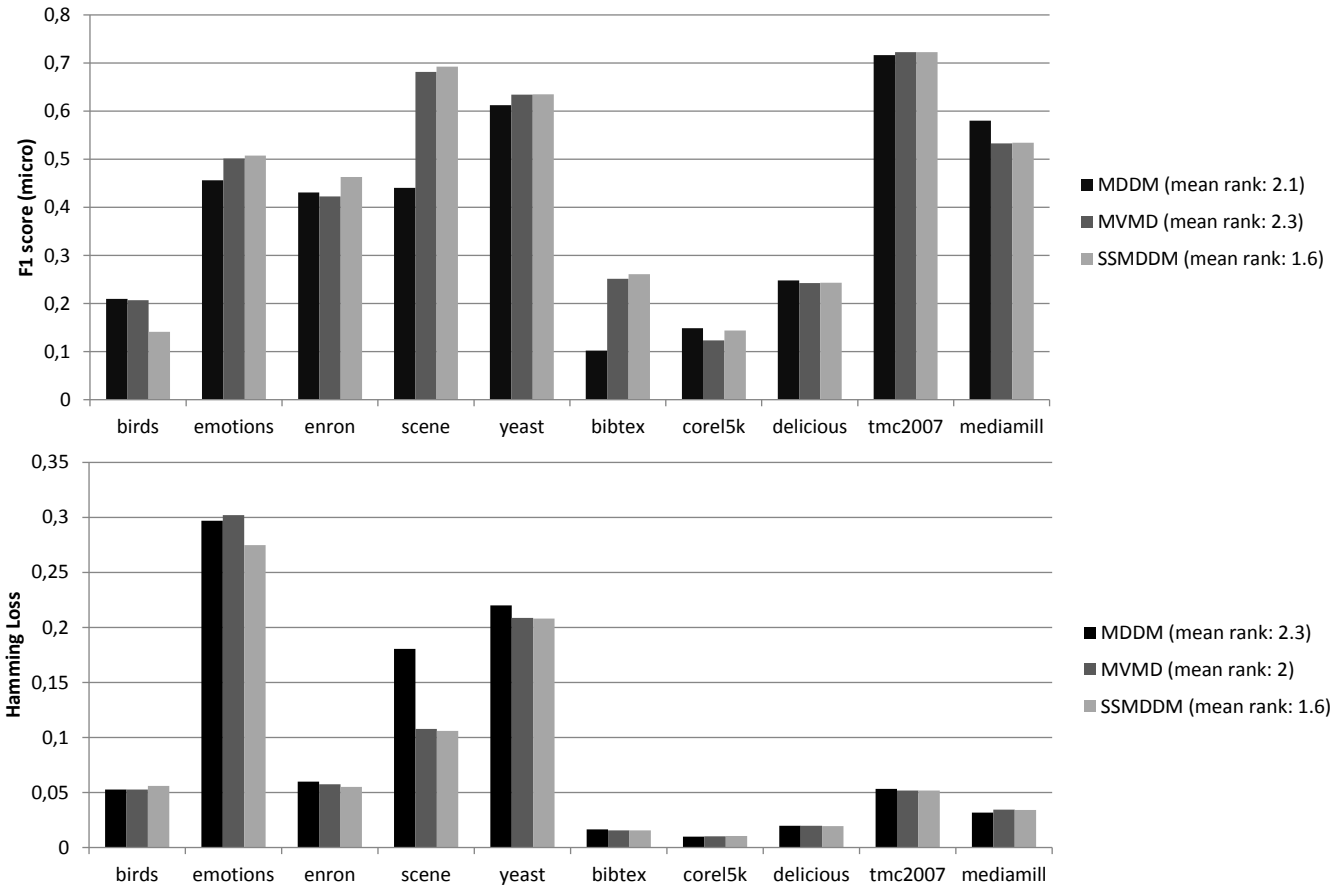

Fig. 4. Comparisons of the dominant algorithms (MDDM, MVMD and SSMDDM) on ten multilabel datasets for F1-score and Hamming Loss.

The dominated methods for the two measures are HSL, DMLDA, MLSI, and CCA. They all belong to C_1 or C_{1-2} due to the lack of Hamming Loss measurements in C_2. They illustrate the bibliographic effect: they are early methods that have been dominated by more recent proposals. More precise interpretations should be made very carefully. HSL performs poorly in the available experiments but it has been considered only once in the corpus $\mathcal{C}$. Moreover, CCA might have been implemented in the majority of papers with a version that is not the optimal one. In the experiments, the feature, label, and feature/label covariance matrices are often badly conditioned due to the multi-label dataset characteristics and it is known that the integration of a slight regularization with a Penrose generalized inversion leads to much better performances [92].

For the set of algorithms from $\mathcal{A}$ which do not belong to these two extremal classes, two cases must be considered. For the algorithms which are not in the G_D graphs, our meta-analysis cannot conclude anything more than the lack of significance of the comparisons which take them into account. For the others, which have both non-null indegree and non-null outdegree, a reasonable recommendation is to replace them with one of the dominant algorithms of their community defined for a similar task.

5 Conclusion

This review is written in a context where multi-label classification is getting a growing attention and meets today's needs to process high dimensional data. To tackle the complexity of the problems, a large number of multi-label dimensionality reduction methods have been published in the last decades. These publications greatly enrich the literature but it remains difficult to link them, to pick the right one for the problem at hand and to determine the work that remains to be done in the topic. Our review attempts to provide these elements. More precisely, we have proposed an overview of the methods through a unifying typology completed by a generalized formulation of the problems and a meta-analysis of the experimental results which can be used as a guideline for algorithm selection and which gives insights for future research.

Overview of the methods: A construction of a typology has been required to disentangle the links between the various methods. It is based on three major criteria. The first one is the space that the methods reduce (feature space, label space or both). Feature space reduction approaches are prevalent for now but with the increasing interest in extreme multi-label learning, methods that also reduce the label space dimensionality are quickly catching up. The second one distinguish the methods which reduce one space by taking into account the information carried by the other from those which perform the reduction independently. Unlike

few years ago, dependent methods predominate today: by preserving the link between the attributes and the labels, they are more efficient for the classification task. The third criterion is the presence/absence of coupling between the classifier and the dimensionality reduction strategy. Today, the two scenarios are very imbalanced and the large majority of the approaches is not coupled with the classifier. In addition to these major structuring aspects, the methods differ in two additional components: the type of transformation (implicit, explicit) that they perform and the constraints that they impose on the problem resolution. Although these differences do not distinguish the approaches on their very nature, they can heavily impact their efficiency and deployability in real-world applications.

Despite all the variability highlighted by the typology, strong similarities between the problems are observable and we have introduced two generic formulations to identify them. The first one scan the problems that rely on loss functions and constraints which can be formulated under a matrix trace minimization and solved by an eigendecomposition. They represent more than half the publications. A more general formulation covers almost the integrality of the publications by integrating the whole set of the implemented ingredients. The combinatorics of these ingredients gives a glimpse of the variety of the approaches. Although a wide spectrum of formulations has been explored, most of the methods focus on ingredients that have interesting properties in terms of numerical optimization (e.g. differentiable, convex, smooth). Then, the chosen resolution method is essential because it affects the scalability (i.e. temporal and spacial complexities) of the reduction algorithm and therefore its potential applications.

Experimental comparisons of the methods: In addition to the theoretical specificities of the approaches, the experimental results remain a major selection criterium. A meta-analysis has been conducted to identify the most significant algorithm performances from the numerical comparisons inventoried in the publications. The results depend on both the used quality measure and the main information guiding the dimensionality reduction (feature space v.s label space or co-label and feature space). Three methods MVMD, SSMDDM and MDDM based on feature space reduction dominate for the two uncorrelated retained measures (Hamming Loss and a selected measure among a large set of correlated ones including Micro F1, Macro F1 and AUC). For the latter, results also highlights SLEEC and REML which are recent approaches especially designed for extreme multi-label learning. A dual examination of domination relationships completes the analysis by pointing out the methods dominated for the two measures. However, from a methodological point of view the generalization of the conclusions should be considered cautiously. As numerous pairwise comparisons are absent of the published experiments, the meta-analysis has been computed on a non-complete graph. Moreover, the heterogeneity of both the datasets used in the different studies and the number of times each algorithm was evaluated add biases to the comparisons. However, despite these limitations, we believe that this first meta-analysis can help identify recurrent properties in the most efficient approaches and also flaws in the experimental protocols (e.g. the lack of some pairwise comparisons). More broadly speaking, the growth of publications in machine learning will certainly foster meta-analysis procedures in the near future.

Insights for future research: The rich literature on dimensionality reduction for multi-label classification offers some major leads of improvement. Theoretical works on stability and robustness guaranties are still at their infancy. In particular, robustness to sampling, to noise, to geometric transformations and to the type of data (e.g. sparse, dense) are major concerns and are almost never addressed. Furthermore, the combinatorics of the key components in the generic formulation could be exploited for future proposals. The coupling between dimensionality reduction and classification especially appears, intuitively and in the experimental comparisons, as a promising component for improving today's state-of-the-art. Finally, the meta-analysis opens a discussion towards the collective construction of a shared experimental protocol which should allow evaluating the performances with limited bias.

ACKNOWLEDGMENTS

The authors would like to thank Doctor Fabrice Clérot, head of the Profiling and Datamining Research Team at Orange Labs (Lannion), for his help on the meta-analysis. They would also like to acknowledge the associate editor and the reviewers whose comments helped us to improve and clarify this manuscript.

REFERENCES

[1] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 518–529, 2011.

[2] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern *et al.*, "The 9th annual mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–8.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.* IEEE, 2009, pp. 248–255.

[4] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for weakly-supervised multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 121–135, 2015.

[5] J.-Y. Jiang, S.-C. Tsai, and S.-J. Lee, "Fsknn: multi-label text categorization based on fuzzy similarity and k nearest neighbors," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2813–2821, 2012.

[6] H. Elghazel, A. Aussem, O. Gharroudi, and W. Saadaoui, "Ensemble multi-label text categorization based on rotation forest and latent semantic indexing," *Expert Systems with Applications*, vol. 57, pp. 1–11, 2016.

[7] X. Wang, W. Zhang, Q. Zhang, and G.-Z. Li, "Multip-schlo: multi-label protein subchloroplast localization prediction with chou's pseudo amino acid composition and a novel multi-label classifier," *Bioinformatics*, vol. 31, no. 16, pp. 2639–2645, 2015.

[8] J.-S. Wu, S.-J. Huang, and Z.-H. Zhou, "Genome-wide protein function prediction through multi-instance multi-label learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 5, pp. 891–902, 2014.

[9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[10] B. Yang, Y. Lei, J. Liu, and W. Li, "Social collaborative filtering by trust," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1633–1647, 2017.

[11] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[12] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[13] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2006.

[14] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[15] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 667–685.

[16] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, 2010.

[17] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

[18] E. Gibaja and S. Ventura, "Multi-label learning: a review of the state of the art and ongoing research," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444, 2014.

[19] Y. Prabhu and M. Varma, "Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 263–272.

[20] J. Weston, A. Makadia, and H. Yee, "Label partitioning for sublinear ranking," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 181–189.

[21] H. Jain, Y. Prabhu, and M. Varma, "Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 935–944.

[22] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[23] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[24] C. J. Burges, "Geometric methods for feature extraction and dimensional reduction-a guided tour," in *Data mining and knowledge discovery handbook*. Springer, 2009, pp. 53–82.

[25] D. J. Hsu, S. M. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," in *Advances in neural information processing systems*, vol. 22, 2009, pp. 772–780.

[26] S. K. Jha and R. Yadava, "Denoising by singular value decomposition and its application to electronic nose data processing," *IEEE Sensors Journal*, vol. 11, no. 1, pp. 35–44, 2011.

[27] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[29] L. Sun, S. Ji, and J. Ye, *Multi-label dimensionality reduction*. CRC Press, 2013.

[30] N. Spolaôr, M. C. Monard, G. Tsoumakas, and H. D. Lee, "A systematic review of multi-label feature selection and a new method based on label construction," *Neurocomputing*, vol. 180, pp. 3–15, 2016.

[31] S. Clinchant and F. Perronnin, "Aggregating continuous word embeddings for information retrieval," in *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, 2013, pp. 100–109.

[32] A. L. Maas and A. Y. Ng, "A probabilistic model for semantic word vectors," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[33] O. Hudry and B. Monjardet, "Consensus theories. an oriented survey," *Mathématiques et sciences humaines. Mathematics and social sciences*, no. 190, pp. 139–167, 2010.

[34] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of the American Control Conference*. IEEE, 2005, pp. 1859–1864.

[35] Y. Zhang and Z.-H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 4, no. 3, p. 14, 2010.

[36] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of NAACL-HLT*, 2013, pp. 746–751.

[37] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in Neural Information Processing Systems*, 2015, pp. 730–738.

[38] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature extraction*. Springer, 2006, pp. 1–25.

[39] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[40] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[41] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[42] A. Blum, "Random projection, margins, kernels, and feature-selection," in *Subspace, Latent Structure and Feature Selection*. Springer, 2006, pp. 52–68.

[43] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[44] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 2008, pp. 90–105.

[45] B. Li, C. Wang, and D.-S. Huang, "Supervised feature extraction based on orthogonal discriminant projection," *Neurocomputing*, vol. 73, no. 1, pp. 191–196, 2009.

[46] L. Sun, S. Ji, and J. Ye, "Hypergraph spectral learning for multi-label classification," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 668–676.

[47] K. Yu, S. Yu, and 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.

[48] J. Xu, J. Liu, J. Yin, and C. Sun, "A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously," *Knowledge-Based Systems*, vol. 98, pp. 172–184, 2016.

[49] F. Tai and H.-T. Lin, "Multilabel classification with principal label space transformation," *Neural Computation*, vol. 24, no. 9, pp. 2508–2542, 2012.

[50] K.-H. Huang and H.-T. Lin, "Cost-sensitive label embedding for multi-label classification," *arXiv preprint arXiv:1603.09048*, 2016.

[51] Y.-N. Chen and H.-T. Lin, "Feature-aware label space dimension reduction for multi-label classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 1529–1537.

[52] J.-J. Zhang, M. Fang, H. Wang, and X. Li, "Dependence maximization based label space dimension reduction for multi-label classification," *Engineering Applications of Artificial Intelligence*, vol. 45, pp. 453–463, 2015.

[53] X. Li and Y. Guo, "Bi-directional representation learning for multi-label classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 209–224.

[54] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the twenty-first international conference on Machine learning (ICML-04)*. ACM, 2004, p. 104.

[55] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[56] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[57] Q. Ye, L. Luo, and Z. Zhang, "Frequent direction algorithms for approximate matrix multiplication with applications in cca," *Computational Complexity*, vol. 1, no. m3, p. 2, 2016.

[58] Y. Zhang and J. Schneider, "Multi-label output codes using canonical correlation analysis," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 873–882.

[59] E. Pacharawongsakda and T. Theeramunkong, "A two-stage dual space reduction framework for multi-label classification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 330–341.

[60] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger, "Supervised semantic indexing," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 187–196.

[61] E. Pacharawongsakda and T. Theeramunkong, "Multi-label classification using dependent and independent dual space reduction," *The Computer Journal*, vol. 56, no. 9, pp. 1113–1135, 2013.

[62] ——, "A comparative study on single and dual space reduction in multi-label classification," in *Knowledge, Information and Creativity Support Systems: Recent Trends, Advances and Solutions*. Springer, 2016, pp. 389–400.

[63] K. Deb, K. Sindhya, and J. Hakanen, "Multi-objective optimization," in *Decision Sciences: Theory and Practice*. CRC Press, 2016, pp. 145–184.

[64] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and multidisciplinary optimization*, vol. 26, no. 6, pp. 369–395, 2004.

[65] I. Giagkiozis and P. J. Fleming, "Pareto front estimation for decision making," *Evolutionary computation*, vol. 22, no. 4, pp. 651–678, 2014.

[66] Y. Guo and D. Schuurmans, "Semi-supervised multi-label classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 355–370.

[67] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proceedings of the 21st international jont conference on Artifical intelligence, IJCAI-09*, 2009, pp. 1077–1082.

[68] T. Mu and S. Ananiadou, "Proximity-based graph embeddings for multi-label classification." in *International Conference on Knowledge Discovery and Information Retrieval*, 2010, pp. 74–84.

[69] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[70] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proceedings of the 21st national conference on Artificial intelligence (AAAI-06)*. AAAI Press, 2006, pp. 421–426.

[71] M. K. Warmuth and D. Kuzmin, "Randomized pca algorithms with regret bounds that are logarithmic in the dimension," in *Advances in neural information processing systems*, 2006, pp. 1481–1488.

[72] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 834–843.

[73] E. Kokiopoulou and Y. Saad, "Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2143–2156, 2007.

[74] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[75] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*. Springer, 2006, pp. 34–51.

[76] S. Ji, L. Tang, S. Yu, and J. Ye, "Extracting shared subspace for multi-label classification," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 381–389.

[77] B. Qian and I. Davidson, "Semi-supervised dimension reduction for multi-label classification," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*. AAAI Press, 2010, pp. 569–574.

[78] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2010, pp. 126–139.

[79] M. Oikonomou and A. Tefas, "Direct multi-label linear discriminant analysis," in *International Conference on Engineering Applications of Neural Networks*. Springer, 2013, pp. 414–423.

[80] P. Li, H. Li, and M. Wu, "Multi-label ensemble based on variable pairwise constraint projection," *Information Sciences*, vol. 222, pp. 269–281, 2013.

[81] G. Luo, T. Huang, and Z. Shi, "Multi-label classification using hypergraph orthonormalized partial least squares," *Journal of Computers*, vol. 9, no. 6, pp. 1364–1370, 2014.

[82] X. Shu, D. Lai, H. Xu, and L. Tao, "Learning shared subspace for multi-label dimensionality reduction via dependence maximization," *Neurocomputing*, vol. 168, pp. 356–364, 2015.

[83] W. Liu and I. W. Tsang, "Large margin metric learning for multi-label prediction," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*. AAAI Press, 2015, pp. 2800–2806.

[84] W. Siblini, R. Alami, F. Meyer, and P. Kuntz, "Supervised feature space reduction for multi-label nearest neighbors," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 182–191.

[85] J. Wicker, B. Pfahringer, and S. Kramer, "Multi-label classification using boolean matrix decomposition," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. ACM, 2012, pp. 179–186.

[86] K. Balasubramanian and G. Lebanon, "The landmark selection method for multiple output prediction," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 283–290.

[87] W. Bi and J. Kwok, "Efficient multi-label classification with many labels," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 405–413.

[88] T. Zhou and D. Tao, "Multi-label subspace ensemble," in *AISTATS*, 2012, pp. 1444–1452.

[89] A. Kapoor, R. Viswanathan, and P. Jain, "Multilabel classification using bayesian compressed sensing," in *Advances in Neural Information Processing Systems*, 2012, pp. 2645–2653.

[90] Z. Lin, G. Ding, M. Hu, and J. Wang, "Multi-label classification via feature-aware implicit label space encoding," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 325–333.

[91] P. Mineiro and N. Karampatziakis, "Fast label embeddings via randomized linear algebra," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 37–51.

[92] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011.

[93] J. Weston, S. Bengio, and N. Usunier, "Wsabie: scaling up to large vocabulary image annotation," in *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI-11)*. AAAI Press, 2011, pp. 2764–2770.

[94] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning (ICML-13)*, 2013, pp. 1247–1255.

[95] F. Mirzazadeh, Y. Guo, and D. Schuurmans, "Convex co-embedding," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI-14)*. AAAI Press, 2014, pp. 1989–1996.

[96] H.-f. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale multi-label learning with missing labels," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 593–601.

[97] P. Rai, C. Hu, R. Henao, and L. Carin, "Large-scale bayesian multi-label learning via topic-based label embeddings," in *Advances in Neural Information Processing Systems*, 2015, pp. 3222–3230.

[98] X. Li and Y. Guo, "Multi-label classification with feature-aware non-linear label space transformation," in *Proceedings of the 24th International Conference on Artificial Intelligence (AAAI-15)*. AAAI Press, 2015, pp. 3635–3642.

[99] C. Xu, D. Tao, and C. Xu, "Robust extreme multi-label learning," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1275–1284.

[100] S. Si, K.-Y. Chiang, C.-J. Hsieh, N. Rao, and I. S. Dhillon, "Goal-directed inductive matrix completion," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1165–1174.

[101] C.-K. Yeh, W.-C. Wu, W.-J. Ko, and Y.-C. F. Wang, "Learning deep latent spaces for multi-label classification," pp. 2838–2844, 2017.

[102] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks*. Springer, 1997, pp. 583–588.

[103] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[104] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[105] M.-L. Zhang and Z.-H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[106] M.-L. Zhang, "Ml-rbf: Rbf neural networks for multi-label learning," *Neural Processing Letters*, vol. 29, no. 2, pp. 61–74, 2009.

[107] Y. Huang, W. Wang, L. Wang, and T. Tan, "Multi-task deep neural network for multi-label learning," in *20th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 2897–2900.

[108] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification—revisiting neural networks," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 437–452.

[109] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218–3229, 2009.

[110] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[111] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[112] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[113] W. Siblini, F. Meyer, and P. Kuntz, "Craftml, an efficient clustering-based random forest for extreme multi-label learning," in *International Conference on Machine Learning*, 2018, pp. 4671–4680.

[114] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[115] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[116] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 2009, pp. 477–485.

[117] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[118] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[119] D. Cai, X. He, and J. Han, "Efficient kernel discriminant analysis via spectral regression," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, 2007, pp. 427–432.

[120] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[121] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik*, vol. 14, no. 5, pp. 403–420, 1970.

[122] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, vol. 35, no. 4, pp. 551–566, 1993.

[123] B. N. Parlett, *The Symmetric Eigenvalue Problem*. SIAM, 1998, vol. 20.

[124] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[125] R. Mises and H. Pollaczek-Geiringer, "Praktische verfahren der gleichungsauflösung." *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, no. 2, pp. 152–164, 1929.

[126] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.

[127] E. Liberty, "Simple and deterministic matrix sketching," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 581–588.

[128] X. Shu, H. Xu, and L. Tao, "A least squares formulation of multi-label linear discriminant analysis," *Neurocomputing*, vol. 156, pp. 221–230, 2015.

[129] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[130] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, 2007.

[131] M. Y. Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.

[132] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[133] D. Zhang, Z.-H. Zhou, and S. Chen, "Semi-supervised dimensionality reduction," in *Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, 2007, pp. 629–634.

[134] P. E. Gill, W. Murray, and M. H. Wright, "Practical optimization," *London: Academic Press, 1981*, 1981.

[135] A. Cauchy, "Méthode générale pour la résolution des systemes d'équations simultanées," *Comp. Rend. Sci. Paris*, vol. 25, no. 1847, pp. 536–538, 1847.

[136] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[137] S. J. Wright, *Primal-dual interior-point methods*. SIAM, 1997.

[138] M. Benzi, "Preconditioning techniques for large linear systems: a survey," *Journal of computational Physics*, vol. 182, no. 2, pp. 418–477, 2002.

[139] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, vol. 2004, 2003.

[140] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer, 2011, pp. 185–212.

[141] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[142] C. Roos, T. Terlaky, and J.-P. Vial, *Theory and algorithms for linear optimization*. Wiley, 1998.

[143] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[144] J. L. Lagrange, *Mécanique analytique*. Mallet-Bachelier, 1853, vol. 1.

[145] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[146] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[147] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007, vol. 382.

[148] G. V. Glass, "Primary, secondary, and meta-analysis of research," *Educational researcher*, vol. 5, no. 10, pp. 3–8, 1976.

[149] W. G. Cochran, "Problems arising in the analysis of a series of similar experiments," *Supplement to the Journal of the Royal Statistical Society*, vol. 4, no. 1, pp. 102–118, 1937.

[150] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in Statistics*. Springer, 1992, pp. 66–70.

[151] K. Ahmad and S. Lily, "The effectiveness of computer applications: A meta-analysis," *Journal of Research on computing in Education*, vol. 27, no. 1, pp. 48–61, 1994.

[152] S. Y. Sohn, "Meta analysis of classification algorithms for pattern recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1137–1144, 1999.

[153] A. Jamain and D. J. Hand, "Mining supervised classification performance studies: A meta-analytic investigation," *Journal of Classification*, vol. 25, no. 1, pp. 87–112, 2008.

[154] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. Merschmann, "Correlation analysis of performance measures for multi-label classification," *Information Processing & Management*, vol. 54, no. 3, pp. 359–369, 2018.
[155] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.
[156] A. B. Kahn, "Topological sorting of large networks," *Communications of the ACM*, vol. 5, no. 11, pp. 558–562, 1962.