Saliency-Based Multilabel Linear Discriminant Analysis

Lei Xu, *Student Member, IEEE*, Jenni Raitoharju, *Member, IEEE*, Alexandros Iosifidis, *Senior Member, IEEE*, and Moncef Gabbouj, *Fellow, IEEE*

***Abstract*—Linear discriminant analysis (LDA) is a classical statistical machine-learning method, which aims to find a linear data transformation increasing class discrimination in an optimal discriminant subspace. Traditional LDA sets assumptions related to the Gaussian class distributions and single-label data annotations. In this article, we propose a new variant of LDA to be used in multilabel classification tasks for dimensionality reduction on original data to enhance the subsequent performance of any multilabel classifier. A probabilistic class saliency estimation approach is introduced for computing saliency-based weights for all instances. We use the weights to redefine the between-class and within-class scatter matrices needed for calculating the projection matrix. We formulate six different variants of the proposed saliency-based multilabel LDA (SMLDA) based on different prior information on the importance of each instance for their class(es) extracted from labels and features. Our experiments show that the proposed SMLDA leads to performance improvements in various multilabel classification problems compared to several competing dimensionality reduction methods.**

***Index Terms*—Class saliency, dimensionality reduction, linear discriminant analysis (LDA), multilabel classification.**

I. Introduction

MULTILABEL classification tasks have become more and more common in the machine-learning field recently, for example, in text information categorization [1], image and video annotation [2], sequential data prediction [3], or music information retrieval [4]. Compared to single-label problems, the characteristics of multilabel problems are more complicated and unpredictable. In a single label problem, each instance merely belongs to a single class. In a multilabel dataset, data items can be associated with either one or several classes. For example, an image can represent both a beach and a sunset and, thus, be associated with both of these classes. Moreover, different classes typically contain a varying number of data items, leading to class-imbalanced problems [5]. Hence, in order to solve a multilabel classification problem efficiently and effectively, we need not only to consider the correlation of class labels and features of each data item but also to take into account the different cardinalities of the classes. The problem of multilabel learning (MLL) has been widely studied and various multilabel classifiers have been suggested [6]–[8].

In this article, we focus on dimensionality reduction for multilabel classification. Dimensionality reduction techniques in general aim at transforming the data to a lower dimensional form that is easier to process by the learning techniques without losing relevant information. The dimensionality reduction techniques for multilabel classification aim at optimizing the data transformation for subsequent multilabel classification. At least 50 such methods have been proposed [9].

A well-known supervised dimensionality reduction technique linear discriminant analysis (LDA) and its variants have been widely used to extract discriminant data representations for solving various problems, for example, in human action recognition [10] or biological data classification [11]. However, they are not optimal for multilabel problems due to the characteristics of multilabel data. This is due to two factors: 1) the contribution of each data item in the calculation of the scatter matrices involved in the optimization problem of single-label LDA and its variants cannot be appropriately determined and 2) the cardinality of the various classes forming the multilabel problem can be quite imbalanced. In multilabel LDA (MLDA) [12] and its variants, these problems have been tackled by introducing different weights to take into account the label and/or feature correlation of different items.

In this article, we propose a novel dimensionality reduction method for multilabel classification based on a probabilistic approach that is able to estimate the contribution of each data item to the classes it is associated with by taking into account prior information encoded using various types of metrics. The proposed calculation of the contribution of each data item to the classes it belongs to can not only weigh its importance but can also avoid problems related to imbalanced classes. To this end, we exploit the concept of class saliency introduced in [13]. Hence, the proposed method is called saliency-based MLDA (SMLDA). Our proposed SMLDA approach exploits

Manuscript received 8 August 2020; revised 8 January 2021; accepted 19 March 2021. Date of publication 20 April 2021; date of current version 16 September 2022. This work was supported by the NSF-Business Finland Center for Visual and Decision Informatics (CVDI) Project AMALIA and EUREKA ITEA 3 Project Mad@Work. The work of Jenni Raitoharju was supported by the Academy of Finland Project under Grant 324475. This article was recommended by Associate Editor B. Ribeiro. *(Corresponding author: Lei Xu.)*

Lei Xu and Moncef Gabbouj are with the Department of Computing Sciences, Tampere University, 33100 Tampere, Finland (e-mail: lei.xu@tuni.fi; moncef.gabbouj@tuni.fi).

Jenni Raitoharju is with the Programme for Environmental Information, Finnish Environment Institute, 40500 Jyväskylä, Finland (e-mail: jenni.raitoharju@syke.fi).

Alexandros Iosifidis is with the Department of Electrical and Computer Engineering, Aarhus University, 8000 Aarhus, Denmark (e-mail: ai@ece.au.dk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCYB.2021.3069338, provided by the authors.

Digital Object Identifier 10.1109/TCYB.2021.3069338

both label and feature information with various prior weighting factors. The proposed method yields features optimized for multilabel classification that can be subsequently classified using any multilabel classifier.

We have made the following contributions on dimensionality reduction for multilabel classification tasks with our novel SMLDA approach.

1) We propose a general framework for using the probabilistic saliency estimation to weigh the importance of each data item for the classes it is associated with for the first time in MLL.
2) We formulate a novel SMLDA method that uses the saliency-based weights in the scatter matrices and can alleviate the problems related to imbalanced datasets.
3) We integrate different label and feature information previously used as weights in dimensionality reduction to SMLDA by using them as prior information for probabilistic saliency estimation and show experimentally that our approach leads to a better performance.
4) We compare our proposed approach to 11 competing dimensionality reduction methods on 17 diverse multilabel datasets using seven evaluation metrics and applying two different multilabel classifiers on the produced features, and the results show considerable improvements in multilabel classification tasks using our approach.

The remainder of this article is structured as follows. In Section II, we briefly review the related works. We include a precise explanation of the LDA and weighted MLDA with adequate mathematical notations to support the derivations of the proposed method. In Section III, we describe our proposed methods in detail. Section IV presents the experimental setup and results. In Section V, we conclude this article and discuss the potential future studies.

II. Related Works

In this section, we first briefly present several dimensionality reduction techniques previously used in multilabel classification tasks in Section II-A. In Section II-B, we give a detailed description of the standard LDA, weighted LDA, and MLDA, since they form the theoretical foundation for the proposed work. Subsequently, we introduce the general concepts of saliency estimation and the probabilistic saliency estimation approach needed to develop the proposed method.

A. Dimensionality Reduction Methods for Multilabel Classification

Dimensionality reduction techniques are commonly used as a preprocessing step for multilabel classification to map the raw high-dimensional data into an optimal lower-dimensional subspace preserving the distinguishing features [9]. The techniques can be categorized as unsupervised or supervised approaches depending on whether class label information is used or not [14]. Furthermore, the techniques can be divided into methods that are independent of the classifiers or dependent of the classifiers [9]. In this article, we consider only dimensionality reduction techniques that are all independent of the classifiers.

Principal component analysis (PCA) [15] is the most well-known unsupervised dimensionality reduction algorithm that minimizes the information lost by preserving as much of the data's variations as possible. Canonical correlation analysis (CCA) [16] is a widely known supervised dimensionality reduction algorithm, projecting the raw data into a subspace exploiting the correlations between the features and labels.

Dimensionality reduction techniques specifically designed for multilabel data include the multilabel-informed latent semantic indexing (MLSI) algorithm [17] that preserves the discriminate feature information by considering the correlations between the multiple labels and multilabel dimensionality reduction via the dependence maximization (MDDM) algorithm [18] that maximizes the dependence between the original features and class labels using the Hilbert–Schmidt independence criterion (HSIC) for measuring dependence. MDDM has two variants with different constraints: 1) MDDMp with an uncorrelated projection constraint and 2) MDDMf with an uncorrelated feature constraint. MDDMp variant was observed to perform better in [18]. Xu *et al.* [19] proposed a multilabel feature extraction method that integrates least-squares formulations of PCA and MDDM linearly, which both maximizes feature variance and maximizes feature-label dependence (MVMD) at the same time.

B. Linear Discrimination Analysis-Based Algorithms for Multilabel Classification

Standard LDA and its variants have been applied to tackle various multilabel classification problems [9], [12], [20]–[23]. These methods operate on N data items $\mathbf{x}_i \in \mathbb{R}^D$ and their corresponding binary label vectors $\mathbf{y}_i \in \{0, 1\}^C$, where D is the original data dimensionality and C is the number of classes. These are arranged into matrices $\mathbf{X} \in \mathbb{R}^{D\times N}$ and $\mathbf{Y} \in \mathbb{R}^{C\times N}$. An element y_{ci} of $\mathbf{Y}$ is 1 only if the corresponding data item $\mathbf{x}_i$ is associated with class c. Thus, in single-label classification tasks, there is a single 1 on each column, but in multilabel classification, the number of 1s is not constrained. The rows of $\mathbf{Y}$ contain 1s for all data items that are associated with the particular class and we denote them as $\mathbf{y}_{(j)}$, where $j \in 1, \ldots, C$. The objective of LDA-based methods is to find a data projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$ that maps the data from the original feature space $\mathbb{R}^D$ to a subspace $\mathbb{R}^d$, where $D > d$, in a manner that maximizes the class discrimination.

1) Linear Discrimination Analysis: LDA is an effective technique to reduce the dimensionality of original data as a prepossessing step for single-label classification problems. LDA operates on within-class, between-class, and total scatter matrices $\mathbf{S}_w$, $\mathbf{S}_b$, and $\mathbf{S}_t$ defined as follows:

$$\mathbf{S}_w = \sum_{c=1}^{C}\sum_{i=1}^{N} y_{ci}\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)^T \tag{1}$$

$$\mathbf{S}_b = \sum_{c=1}^{C}\left(\sum_{i=1}^{N} y_{ci}\right)\big(\boldsymbol{\mu}_c - \boldsymbol{\mu}\big)\big(\boldsymbol{\mu}_c - \boldsymbol{\mu}\big)^T \tag{2}$$

$$\mathbf{S}_t = \sum_{c=1}^{C}\sum_{i=1}^{N} y_{ci}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{3}$$

Here, $\boldsymbol{\mu}_c$ denotes the mean vector of class c as

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N} y_{ci}\mathbf{x}_i \tag{4}$$

where $N_c = \sum_{i=1}^{N} y_{ci}$ is the cardinality of class c. The total mean vector $\boldsymbol{\mu}$ is computed as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i. \tag{5}$$

The optimal projection matrix $\mathbf{W}$ is learned by maximizing Fisher's discriminant criterion [24] that minimizes the within-class scatter while maximizing the between-class scatter

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)} \tag{6}$$

where tr(.) denotes the trace of a matrix. Typically, the solution to this trace ratio optimization is approximated by solving the corresponding ratio trace optimization. This allows obtaining the projection matrix $\mathbf{W}$ by solving the generalized eigenvalue problem

$$\mathbf{S}_b\mathbf{w} = \mathbf{S}_w\lambda\mathbf{w} \tag{7}$$

and taking the eigenvectors corresponding to the $d \leq C-1$ largest eigenvalues as columns of the projection matrix $\mathbf{W}$. The rank of $\mathbf{S}_b$ is equal to $C-1$, which is the maximal dimensionality of the resulting subspace. Also, different iterative methods for solving directly the trace ratio problem have been proposed [25], [26].

Since $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$, an alternative approach is to use $\mathbf{S}_t$ instead of $\mathbf{S}_w$ and maximize Fisher's discriminant criterion as

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_t\mathbf{W}\right)}. \tag{8}$$

Finally, the optimized features can be obtained as

$$\mathbf{Z} = \mathbf{W}^T\mathbf{X}. \tag{9}$$

The datasets used in most traditional LDA classification tasks are assumed to have equal class distribution as a homoscedastic Gaussian model [27], in which the covariance matrices of each class should be identical [28]. The performance is affected severely due to the imbalance of input datasets [29].

2) Weighted Linear Discrimination Analysis: In order to enhance the robustness of traditional LDA on different kinds of datasets, various weight factors based on class statistics [26], [28], [30], for example, class cardinality, a prior probability, have been introduced into the definitions of the scatter matrices to balance the contribution of each class. Weighted LDA approaches have diminished the influence of outlier classes on the scatter matrices of imbalanced datasets to some extent; however, they still neglect the varying importance of individual samples in the class description. Saliency-based weighted LDA (SwLDA) [13] was proposed to explore the contribution of each instance based on probabilistic saliency estimation [31]. Our work uses a similar idea for multilabel classification.

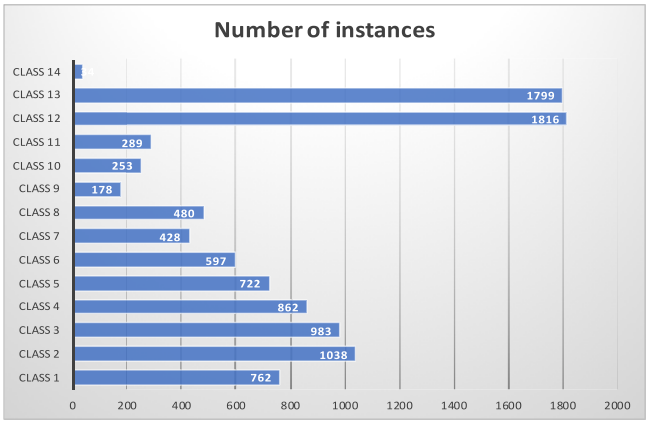

Fig. 1. Number of instances for each class in the Yeast database.

3) Multilabel Linear Discrimination Analysis: Although weighted LDA algorithms enhance the performance in single-label classification tasks [32] compared to traditional LDA, such variants are still not directly applicable for multilabel classification tasks [12]. In a multilabel dataset, label information contains certain correlations or dependencies [33], for example, an image instance labeled as "car" highly correlates to label "road" [12]. Besides, it is quite common that the number of samples in each class in a multiclass dataset is imbalanced. For example, the largest class size is 1128 and the smallest is 21 in the widely used Yeast database [34], as shown in Fig. 1. Due to the specific characteristics of multilabel databases, it is imperative to take into account the correlation of class labels and/or discriminative feature information of each instance to tackle the suboptimal classification result on imbalanced datasets.

If traditional LDA and its variants are applied to multilabel classification tasks by simply using (1) and (2) with the multilabel label matrix $\mathbf{Y}$, an overcounting problem is encountered, that is, the contribution of one instance can be repeatedly counted in computing the scatter matrices. Hence, an MLDA [12] and its variants use weight factors to express redundancy or/and correlation information so that the scatter matrices can be calculated without redundancy on multilabel databases. These weight factors can be organized to a non-negative weight matrix $\mathbf{M} \in \mathbb{R}^{C\times N}$ with the same size as the label matrix $\mathbf{Y}$

$$\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_N] = \left[\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(j)}, \ldots, \mathbf{m}_{(C)}\right]^T \tag{10}$$

where $\mathbf{m}_i$ represents a weight vector for the ith instance, $\mathbf{m}_{(j)}$ is a weight vector for the jth class, and m_{ci} is the weight factor of the ith instance for class c.

We denote by n_i, $n_{(c)}$, and n the summations of the weights for the ith instance, weights for class c, and all weights, respectively

$$n_i = \sum_{c=1}^{C} m_{ci} \tag{11}$$

$$n_{(c)} = \sum_{i=1}^{N} m_{ci} \tag{12}$$

$$n = \sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci} = \sum_{c=1}^{C} n_{(c)}. \tag{13}$$

We also define row vectors $\hat{\mathbf{n}}$ and $\hat{\mathbf{m}}$ and matrix $\hat{\mathbf{M}}$ for simplifying notations as

$$\hat{\mathbf{n}} = \left[\frac{1}{n_{(1)}}, \ldots, \frac{1}{n_{(C)}}\right]. \tag{14}$$

$$\hat{\mathbf{m}} = [n_1, \ldots, n_i, \ldots, n_N] = \sum_{c=1}^{C} \mathbf{m}_{(c)} \tag{15}$$

$$\hat{\mathbf{M}} = \mathbf{M}\mathrm{diag}\left(\hat{\mathbf{n}}^{\frac{1}{2}}\right) \tag{16}$$

where $\hat{\mathbf{M}}$ has row vectors $([\mathbf{m}_{(c)}]/[\sqrt{n_{(c)}}])$ for $c = 1, \ldots, C$.

The scatter matrices for MLDA can now be given as

$$\begin{aligned}\mathbf{S}_w &= \sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}(\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T \\ &= \mathbf{X}\left(\mathrm{diag}(\hat{\mathbf{m}}) - \hat{\mathbf{M}}^{\intercal}\hat{\mathbf{M}}\right)\mathbf{X}^{\intercal} \end{aligned}\tag{17}$$

$$\begin{aligned}\mathbf{S}_b &= \sum_{c=1}^{C}\left(\sum_{i=1}^{N} m_{ci}\right)(\boldsymbol{\mu} - \boldsymbol{\mu}_c)(\boldsymbol{\mu} - \boldsymbol{\mu}_c)^T \\ &= \mathbf{X}\left(\hat{\mathbf{M}}^{\intercal}\hat{\mathbf{M}} - \frac{1}{n}\hat{\mathbf{m}}^{\intercal}\hat{\mathbf{m}}\right)\mathbf{X}^{\intercal} \end{aligned}\tag{18}$$

$$\begin{aligned}\mathbf{S}_t &= \mathbf{S}_w + \mathbf{S}_b \\ &= \mathbf{X}\left(\mathrm{diag}(\hat{\mathbf{m}}) - \frac{1}{n}\hat{\mathbf{m}}^{\intercal}\hat{\mathbf{m}}\right)\mathbf{X}^{\intercal} \end{aligned}\tag{19}$$

where $\boldsymbol{\mu}$ is the total mean vector of all training instances and $\boldsymbol{\mu}_c$ is the mean vector of class c

$$\boldsymbol{\mu} = \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}\mathbf{x}_i}{\sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}}, \quad \boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} m_{ci}\mathbf{x}_i}{\sum_{i=1}^{N} m_{ci}}. \tag{20}$$

A detailed derivation of the matrix forms in (17)–(19) can be found in [14]. The optimal projection matrix $\mathbf{W}$ can still be obtained by solving the generalized eigenproblem in (7) as discussed in Section II.

In the original MLDA [12], the weight factors are solved using label correlations for different classes. First, a correlation matrix $\mathbf{R} \in \mathbb{R}^{C\times C}$ is computed using the class labels of each pair of classes

$$R_{kl} = \cos(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}) = \frac{\mathbf{y}_{(k)}^T\mathbf{y}_{(l)}}{\|\mathbf{y}_{(k)}\|\|\mathbf{y}_{(l)}\|} \tag{21}$$

where $\mathbf{y}_{(k)}$, $\mathbf{y}_{(l)}$ are label vectors for classes $k, l \in 1, \ldots, C$. The label correlation for classes k and l is high if the classes are closely related. The correlation matrix $\mathbf{R}$ can be used to compute the weight matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{RY}$. However, also this approach may lead to the overcounting problem. To tackle the overcounting problem [12], the weight factors are normalized with the ℓ_1-norm

$$\mathbf{m}'_i = \frac{\mathbf{m}_i}{\|\mathbf{y}_i\|_{\ell_1}}. \tag{22}$$

Other metrics for evaluating the relationships among instances from the labels and/or features were used for determining the weights in [14] under the name weighted multilabel LDA (wMLDA). In addition to the label correlation-based weight factors used in MLDA [12], Xu [14] considered entropy-based [35], binary-based [20], fuzzy-based [36], and dependence-based weight factors [14]. Similar metrics can be used as prior information within our probabilistic saliency estimation framework. Therefore, the detailed explanations of these metrics are left to Section III-A1.

In [21], MLDA was extended to Direct MLDA by changing the definition of $\mathbf{S}_b$ in a way that allows obtaining a higher dimensional subspace than the original MLDA, where the subspace dimensionality is limited by the rank of $\mathbf{S}_b$ to $C-1$. This extension work further enhanced the results in multilabel video classification tasks. Another extension, multilabel discriminant analysis with locality consistency (MLDA-LC) [22] not only preserves the global class label information as MLDA does but also incorporates a graph regularized term to utilize the local geometric information. MLDA-LC reveals the similarity among nearby instances with transformation in the projection space using incorporation of the graph Laplacian matrix into the MLDA approach, which further enhances the classification performance in multilabel datasets compared to MLDA and MLLS algorithms.

C. Saliency Estimation

Saliency estimation, as a standard computer vision task, is inspired by neurobiological studies [37] and cognition psychology [38]. Generally, saliency estimation is a preprocessing step for various high-level computer vision tasks, such as object detection [31], [39] and omni directional images [40]. Saliency in physiological science is defined as a special kind of perception of the human visual system, by which humans can perceive particular parts in a scene in details due to colors, textures, or other prominent information contained in these parts [41]. These particular parts can be distinguished as a foreground from nonsalient background parts.

Computational saliency estimation approaches can be categorized as local approaches and global approaches based on the way they process saliency information [41]. Local saliency estimation approaches explore the prominent information around the neighborhood of specific pixels/regions whilst global approaches exploit the rarity of a pixel/patch/region in the entire scene. Since the emergence of the computational saliency estimation field [42], various probabilistic approaches have been explored in this topic.

Aytekin *et al.* [31] proposed a probabilistic saliency estimation approach for segmenting salient objects in an image, where a probability mass function $P(\mathbf{x})$ depicts whether a region $\mathbf{x}_i$ (pixel, super-pixel, or patch) in an image is considered as a distinct region. The higher the values of $P(\mathbf{x}_i)$ for a region, the more prominent the region is. $P(\mathbf{x})$ is solved by simultaneously optimizing two terms to allocate not only lower probabilities to nonsalient regions but also similar probabilities to similar regions

$$\begin{aligned}&\underset{P(x)}{\mathrm{argmin}} \left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\left(P(\mathbf{x}_i) - P(\mathbf{x}_j)\right)^2 a_{ij}\right) \\ &\quad = \underset{P(x)}{\mathrm{argmin}}\left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\left(P(\mathbf{x}_i)^2 - P(\mathbf{x}_i)P(\mathbf{x}_j)\right)a_{ij}\right) \\ &\quad \text{s.t.} \sum_i P(\mathbf{x}_i) = 1 \end{aligned}\tag{23}$$

where the first term suppresses the probability of a nonprominent region $\mathbf{x}_i$ using its prior information $v_i \geq 0$. In the second term, a high similarity of regions $\mathbf{x}_i$ and $\mathbf{x}_j$, given as a high similarity value a_{ij}, forces the regions to have similar probabilities. To go from the first form to the second form, the similarity values are assumed symmetric, that is, $a_{ij} = a_{ji}$.

The optimization task in (23) can be expressed using matrix notations as

$$\begin{aligned} \mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \ & \left(\mathbf{p^T H p}\right) \\ & \mathbf{H} = \mathbf{D} - \mathbf{A} + \mathbf{V} \\ \text{s.t.} \ & \mathbf{p}^T\mathbf{1} = 1 \end{aligned} \tag{24}$$

where $\mathbf{p}$ is a probability vector that contains the probabilities of each element or region $\mathbf{x}_i$ to be salient, that is, $p_i = P(\mathbf{x}_i)$, $\mathbf{A}$ is an affinity matrix, which denotes the similarity of each pair of regions $\mathbf{x}_i$ and $\mathbf{x}_j$ as $[\mathbf{A}]_{ij} = a_{ij}$. $\mathbf{D}$ is a diagonal matrix having elements equal to $[\mathbf{D}_{ii}] = \sum_j a_{ij}$, $\mathbf{V}$ is a diagonal prior information matrix having elements $[\mathbf{V}]_{ii} = v_i$, and $\mathbf{1}$ is a vector of ones. Then, the Lagrangian multiplier method is employed

$$\mathcal{L}(\mathbf{p}, \gamma) = \left(\mathbf{p^T H p}\right) - \gamma\left(\mathbf{p^T 1} - 1\right). \tag{25}$$

A global optimum $\mathbf{p}^*$ is obtained by setting the partial derivative of (25) with the respect $\mathbf{p}$ to 0. The final optimized probability vector is

$$\mathbf{p}^*_{pse} = \frac{1}{\mathbf{1}^T\mathbf{H}^{-1}\mathbf{1}}\mathbf{H}^{-1}\mathbf{1} \tag{26}$$

where the normalization constant $\mathbf{1}^T\mathbf{H}^{-1}\mathbf{1}$ follows from the constraint $\mathbf{p}^T\mathbf{1} = 1$ and ensures that the resulting values are actual probabilities. Due to the properties of matrix $\mathbf{H}^{-1}$, the elements of $\mathbf{p}^*$ are always non-negative as shown in [31]. A more detailed derivation of (26) can be also found in [31].

III. Proposed Method

We propose a novel SwLDA method for multilabel classification tasks. The proposed method has two main steps. For the first step, we propose a probabilistic saliency estimation approach to evaluate the importance of each sample for each class in a multilabel dataset. This is a general framework for multilabel class-saliency and, as future work, can be easily integrated also with other dimensionality reduction techniques or directly with multilabel classifiers that weigh the samples based on their importance. In the second step, we use the class-saliency analysis as weights in an MLDA technique.

In our prior work [13], we used the idea of probabilistic class-saliency estimation for single-label datasets to tackle the suboptimal results of LDA-based algorithms caused by imbalanced datasets or/and outliers. In this article, we formulate multilabel extensions of both the probabilistic class-saliency estimation and the subsequent LDA-based dimensionality reduction technique. Furthermore, we show how to use as prior information in the probabilistic multilabel class-saliency estimation framework different types of information extracted from the data and/or labels that have been previously used directly as sample weights in MLL and we propose a new misclassification-based multilabel information extraction approach, which is based on the prior information type used for single-label data in [13].

A. Probabilistic Multilabel Class-Saliency Estimation

The goal of probabilistic multilabel class-saliency estimation is to define the probability of each data item to be salient for each class. In other words, we want to find a probability matrix $\mathbf{P} \in \mathbb{R}^{C\times N}$

$$\mathbf{P} = \left[\mathbf{p}_1, \ldots, \mathbf{p}_i, \ldots, \mathbf{p}_N\right] = \left[\mathbf{p}_{(1)}, \ldots, \mathbf{p}_{(j)}, \ldots, \mathbf{p}_{(C)}\right]^T \tag{27}$$

where $\mathbf{p}_i \in \mathbb{R}^C$ is a vector containing the probabilities for instance i to be salient for class c and $\mathbf{p}_{(j)} \in \mathbb{R}^N$ is the probability vector for the jth class. The probabilities for each class are normalized to sum up to one, that is, $\sum_{i=1}^{N} p_{ci} = 1 \ \forall c \in 1, \ldots C$.

First, we make an assumption that only data items associated with a class can be salient, that is, $p_{ci} = 0$ if $y_{ci} = 0$. As we need to solve the probabilities p_{ci} only for data items associated with class c, we form a reduced data matrix $\mathbf{X}^c \in \mathbb{R}^{D\times N^c}$ and reduced probability vector $\mathbf{p}^c \in \mathbb{R}^{N^c}$ corresponding to N^c data items associated with class c. Now, we can write the optimization problem of probabilistic multilabel class-saliency estimation as

$$\begin{aligned} \underset{\mathbf{p}^c}{\operatorname{argmin}} \ & \left(\sum_i^{N^c}\left(p_i^c\right)^2 v_i^c + \frac{1}{2}\sum_i^{N^c}\sum_j^{N^c}\left(p_i^c - p_j^c\right)^2 a_{ij}^c\right) \\ & = \underset{\mathbf{p}^c}{\operatorname{argmin}}\left(\sum_i^{N^c}(p_i^c)^2 v_i^c + \frac{1}{2}\sum_i^{N^c}\sum_j^{N^c}\left((p_i^c)^2 a_{ij}^c + (p_j^c)^2 a_{ij}^c\right)\right. \\ & \qquad \left. - \sum_i^{N^c}\sum_j^{N^c}\left(p_i^c p_j^c\right)a_{ij}^c\right) \\ \text{s.t.} \ & \sum_i^{N^c} p_i^c = 1 \end{aligned} \tag{28}$$

where p_i^c is the ith element in $\mathbf{p}^c$ and $v_i^c \geq 0$ is the corresponding prior information to suppress the probabilities of nonsalient instances from class c. The similarity value a_{ij}^c forces the instances $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ have similar probabilities, if they are similar. Unlike the original probabilistic saliency estimation in (23), we do not require the similarity values to be symmetric.

Equation (28) can be expressed in matrix notation as

$$\begin{aligned} \mathbf{p^{c*}} = \underset{\mathbf{p}^c}{\operatorname{argmin}} \ & \left(\mathbf{p}^{cT}\mathbf{H}^c\mathbf{p}^c\right) \\ & \mathbf{H}^c = \frac{1}{2}\mathbf{D_1}^c + \frac{1}{2}\mathbf{D_2}^c - \mathbf{A}^c + \mathbf{V}^c \\ \text{s.t.} \ & \mathbf{p}^{cT}\mathbf{1} = 1 \end{aligned} \tag{29}$$

where $\mathbf{A}^c$ is an affinity matrix of the items associated with class c with $[\mathbf{A}^c]_{ij} = a_{ij}^c$ expressing the similarity of the ith and jth class items, the diagonal matrix $\mathbf{D_1}^c$ can be then computed as $[\mathbf{D_1}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ij}$ and $\mathbf{D_2}^c$ can be then computed as $[\mathbf{D_2}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ji}$, that is, $\mathbf{D_1}^c$ has summations over rows

while $\mathbf{D_2}^c$ has summations over columns. $\mathbf{V}^c$ is a diagonal prior information matrix having elements $[\mathbf{V}^c]_{ii} = v_i^c$.

In this work, the compute the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N_c \times N_c}$ with the RBF kernel function as

$$\left[\mathbf{A}^c\right]_{ij} = \exp\left(-\frac{\left\|\mathbf{x}_i^c - \mathbf{x}_j^c\right\|^2}{2\sigma^2}\right) \tag{30}$$

where $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ are the ith and jth instance in class c and σ is a hyper parameter. While (30) is sensitive to the parameter σ, we follow a common approach of setting its values to the mean distance value between the training samples. The affinity matrix could be also replaced by sparse variants, for example, by forming a kNN graph and keeping only the values for k nearest neighbors or by using an affinity matrix proposed in [43], where the sensitive parameter σ is avoided.

$\mathbf{V}^c \in \mathbb{R}^{N^c \times N^c}$ is a diagonal matrix, which carries the prior information on whether each instance in class c is salient for the class. The values of $\mathbf{V}^c$ are higher for samples, which are expected to *not* be salient, that is, the lower a value $[\mathbf{V}^c]_{ii}$, the more prominent the corresponding ith instance is expected to be. Values for $[\mathbf{V}^c]_{ii} = v_i^c \ \forall i \in 1, \ldots N_c$ can be estimated from different prior information. For example, a data item that belongs to all the classes it is unlikely to be salient for any particular class or if an item is very different from other samples in a class it is unlikely to be salient for that class. It should be noted that while we set prior information values v_i^c only for items associated with class c, we can exploit the information extracted from other data items while setting the values of $\mathbf{V}^c$. For example, items having a high similarity with many items not associated with the class could be considered less likely to be prominent. We introduce six different approaches to set the values of $\mathbf{V}^c$ in Section III-A1.

After computing the matrices $\mathbf{A}^c$ and $\mathbf{V}^c$, the probability vector $\mathbf{p^{c*}}$ can be solved as

$$\mathbf{p}^{c*} = \frac{1}{\mathbf{1}^T\mathbf{H}^{c-1}\mathbf{1}}\mathbf{H}^{\mathbf{c}-1}\mathbf{1}. \tag{31}$$

In order to avoid singularity during this process, a regularized version of $\mathbf{H}^c$ with a small value ϵ added to the diagonal elements if $\mathbf{H}^c$ is rank-deficient.

As the probability vector $\mathbf{p}^{c*}$ obtained by solving (31) has only N^c elements, but we want to form a probability matrix $\mathbf{P} \in \mathbb{R}^{C \times N}$ shown in (27), we need to put the values $\mathbf{p}^{c*}$ to the correct places in $\mathbf{P}$. If the ith item in class c is the jth item in the entire dataset, this can be done by setting $[\mathbf{P}]_{cj} = p_i^c$ for all items in class c. To obtain full matrix $\mathbf{P}$, the above-described process is repeated for each class $c \in \{1, \ldots, C\}$. The sum of the values for each row in $\mathbf{P}$ is one, which is expected to alleviate the overcounting problem.

1) Prior Information Types: Probabilistic saliency estimation [31] was originally proposed for segmenting salient parts from images. In this setup, the prior information used was that the pixel at the image borders is typically nonsalient. The prior information value v_i was set to 1 for any border pixels and to 0 for all the others. In multilabel class-saliency estimation, we similarly want to use v_i^c to integrate our prior knowledge on which data items are likely to be salient for class c. To this end, we propose a novel information type for MLL context: misclassification-based prior information. Furthermore, we introduce five prior information types based on weight factors proposed for MLDA and wMLDA. Our experimental results show that using these information types as prior information for our proposed saliency estimation framework instead of using them directly as weight factors consistently leads to better results.

Correlation-based prior information (SMLDAc) was used as weight factors in the original MLDA algorithm [12]. As in [12], we first compute the label correlation matrix $\mathbf{R}$ defined in (21). We then compute the normalized weight vector $\mathbf{m}'_j \in \mathbb{R}^C$ using (22) for all data items and set our prior information matrix values as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - m'_{cj} \tag{32}$$

where item j of the full dataset is the ith item associated with class c. Label correlation information is widely exploited to tackle the redundancy of label information in multilabel tasks [12], [44], but it can lead to a suboptimal result due to nonzero values in the correlation weight factor matrix for irrelevant labels [14]. As we pick only the values for data items associated with class c, the problem of unwanted nonzero values can be avoided.

Binary-based prior information (SMLDAb) utilizes the label information as in [20]. In our formulation, this approach reduces to having an equal value in $\mathbf{V}^c$ for all instances as only instances belonging to class c are considered in $\mathbf{V}^c$. For wMLDA, such direct use of class labels leads to an overcounting problem in the scatter matrices. In our formulation, this problem is avoided because $\mathbf{V}^c$ merely represents the prior information for class saliency estimation and the final weight matrix $\mathbf{P}$ is normalized for each class.

Entropy-based prior information (SMLDAe) assumes that data items, which are associated with more classes are less salient for any class as in [14] and [35]. We use this assumption as our prior information as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - \frac{1}{\left\|\mathbf{y}_i^c\right\|_{\ell_1}} \tag{33}$$

where $\mathbf{y}_i^c$ is the label vector of the ith sample associated with class c and, thus, $\|\mathbf{y}_i^c\|_{\ell_1}$ is the total number of classes the item is associated with.

Fuzzy-based prior information (SMLDAf) uses a supervised version of fuzzy C-means clustering algorithm (SFCM) as in [14] and [36] to learn the membership degree of each item in each class. We use the membership directly as our prior information as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - g_j^c \tag{34}$$

where g_j^c is the membership degree of item j in class j and item j is the ith item associated with class c.

Dependence-based prior information (SMLDAd) uses HSIC [45], which is used to describe statistical dependence between features and labels based on the estimation of the Hilbert–Schmidt norms. To maximize HSIC, we follow an iterative algorithm described in [14].

This approach transforms a multilabel task to several single-label tasks. It allocates 1 to only one prominent class for each item after the final iteration. In our probabilistic formulation, we set

$$\left[\mathbf{V}^c\right]_{ii} = 1 - h_j^c \tag{35}$$

where h_j^c is 1 if item j has been assigned to class c and 0 otherwise and item j is the ith item associated with class c.

Misclassification-based prior information (SMLDAm) is similar to the prior information used in [13] for single-label data to alleviate the suboptimal result in LDA arising from outlier items on imbalanced datasets

$$[\mathbf{V}^c]_{ii} = \begin{cases} 0, & \text{if } d_{ic}^c < \min\limits_{k \neq c} d_{ic}^k \\ \dfrac{d_{ic}^c}{\min\limits_{k \neq c} d_{ic}^k}, & \text{otherwise} \end{cases} \tag{36}$$

where $d_{ic}^k = \|\mathbf{x}_{ic} - \boldsymbol{\mu}_k\|_2^2$, $\mathbf{x}_{ic}$ is the ith instance of class c, and $\boldsymbol{\mu}_k$ is the mean vector of class k. Using this prior information type, a sample that is closer to another class is considered less salient for class c even if it is relatively close to the center of class c. Note that when computing this prior information matrix, we consider the full data $\mathbf{X}$ and not only the data items in $\mathbf{X^c}$ for which we are defining the prior information values.

B. Saliency-Based Multilabel Linear Discriminant Analysis

After forming the probability matrix $\mathbf{P}$ using the proposed probabilistic multilabel class-saliency estimation, we use the probabilities directly as weights for our MLDA. We compute the scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ as

$$\mathbf{S}_w = \mathbf{X}\big(\text{diag}\big(\hat{\mathbf{p}}\big) - \mathbf{P}^\intercal\mathbf{P}\big)\mathbf{X}^\intercal \tag{37}$$

$$\mathbf{S}_b = \mathbf{X}\left(\mathbf{P}^\intercal\,\mathbf{P} - \frac{1}{n}\hat{\mathbf{p}}^\intercal\hat{\mathbf{p}}\right)\mathbf{X}^\intercal \tag{38}$$

where $\hat{\mathbf{p}} = \sum_{c=1}^{C} \mathbf{p}_{(c)}$ and $n = \sum_{c=1}^{C}\sum_{i=1}^{N} p_{ci}$. Note that the probability values for each class are always normalized to sum to one. By setting $m_{ci} = p_{ci}$, we get $n_{(c)} = 1$ from (12) for all classes and, thus, $\hat{\mathbf{n}}$ in (14) is a vector of ones and $\text{diag}(\hat{\mathbf{n}}^{[1/2]})$ in (16) is an identity matrix. This gives us simpler formulas for $\mathbf{S}_w$ and $\mathbf{S}_b$ than the ones used in MLDA.

The optimal projection matrix $\mathbf{W}$ can be obtained by solving the regularized version of the generalized eigenproblem in (7)

$$\mathbf{S}_b\mathbf{w} = (\mathbf{S}_w + \epsilon\mathbf{I})\lambda\mathbf{w} \tag{39}$$

where ϵ is a small constant added to the diagonal values of $\mathbf{S}_w$ to avoid problems caused by singularity. We select the eigenvectors corresponding to d largest eigenvalues containing 0.999 of the information to form the projection matrix $\mathbf{W}$ and, finally, the features optimized for multilabel classification can be obtained as

$$\mathbf{Z} = \mathbf{W}^T\mathbf{X}. \tag{40}$$

The pseudocode for the overall SMLDA algorithm is provided in Algorithm 1. In the pseudocode, we give the correlation-based prior information type as our default type, but other prior information types can be used by simply replacing (32) on the pseudocode line 4 with a formula of another prior information type.

Algorithm 1: The Pseudocode of SMLDA

```
/* Training procedure for obtaining
   optimal projection matrix W        */
```

Input: $\mathbf{X}_{train} \in \mathbb{R}^{D\times N}$, $\mathbf{Y}_{train} \in \mathbb{R}^{C\times N}$
Output: Projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$

1 Create the probability matrix $\mathbf{P} \in \mathbb{R}^{C\times N}$ and fill it with zeros;
2 **for** *each class* $c \in \{1, \ldots, C\}$ **do**
3 Calculate the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N^c\times N^c}$ using (30);
4 Calculate the prior information matrix $\mathbf{V}^c \in \mathbb{R}^{N^c\times N^c}$ using (32);
5 Calculate diagonal matrices $\mathbf{D_1}^c, \mathbf{D_2}^c \in \mathbb{R}^{N^c\times N^c}$ as $[\mathbf{D_1}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ij}$ and $[\mathbf{D_2}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ji}$;
6 Calculate $\mathbf{H}^c = \frac{1}{2}\mathbf{D_1}^c + \frac{1}{2}\mathbf{D_2}^c - \mathbf{A}^c + \mathbf{V}^c$;
7 Using Eq. (31), solve the probability matrix $\mathbf{p}^{c*}$;
8 Put the values of $\mathbf{p}^{c*}$ to correct places in $\mathbf{P}$;
9 **end**
10 Calculate the scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ using Eqs. (37) and (38);
11 Solve the projection matrix $\mathbf{W}$ using Eq. (39);

C. Computational Complexity Analysis

The computational complexity of the proposed SMLDA algorithm is formed as follows: for a class with N^c associated data items, the computational complexity of computing the kernel matrix is $([N^{c2} - N^c]/2)$, that is, the complexity of computing the affinity matrix is $\mathcal{O}(N^{c2})$. The complexity of computing the prior information matrix using (32) is $\mathcal{O}(C^2N)$ as it requires computing the correlation between each pair of classes using (21) and multiplying $C \times C$ and $C \times N$ matrices in (22). The computational complexity of solving (31) is $\mathcal{O}(N^{c3})$ due to the required matrix inversion. The overall complexity of applying the probabilistic multilabel class-saliency estimation for all the classes becomes $\mathcal{O}(\max_c N^{c3})$. The complexity of the LDA operation for D-dimensional data items is $\mathcal{O}(D^3)$. The overall complexity of SMLDA is $\mathcal{O}(\max_c N^{c3} + D^3)$. Thus, if $\max_c N^c < D$, the proposed method does not significantly affect the complexity compared to the standard LDA operation, but for $\max N^c > D$ the complexity is higher.

IV. Experiments

A. Databases and Data Preprocessing

We performed our experiments on 17 publicly available multilabel databases[1,2]. The datasets and their characteristics are given in Table I, where "Cardinality" means the mean numbers of class labels per instance for the training set and "Min #/Max #" shows the smallest/largest class size in the training set. The mean imbalance ratio ("MeanIR") measures the dataset imbalance following [59], where the imbalance for a class is computed by dividing the largest class size by the

[1] http://ceai.njnu.edu.cn/Lab/LABIC/LABIC_Software.html
[2] http://www.uco.es/kdis/mllresources/#KatakisEtAl2008

TABLE I
CHARACTERISTICS OF DATASETS USED FOR EXPERIMENTS

Database	Contents	Train #	Test #	Classes	Features	Cardinality	Min #	Max #	meanIR	meanCIR
Bibtex [46]	Text	4880	2515	159	1836	2.4	28	691	12.8	89.3
Birds [47]	Audio	179	172	19	260	1.9	4	64	6.1	16.1
Cal500 [48]	Music	300	202	174/173	68	26.1	2	263	21.1	23.1
CHD_49 [49]	Medicine	371	181	6	49	2.6	12	281	5.3	6.6
Corel16k(001) [50], [14]	Scene	5188	1744	153	500	3.1	21	1124	23.8	108.8
Emotions [51]	Music	398	195	6	72	1.9	96	181	1.5	2.4
Enron [52]	Text	988	660	57/53	1001	27	0	535	74.8	137.1
Eukaryote [19]	Biology	4658	3108	22	440	1.1	6	1387	45.1	150.5
Human [53]	Biology	1862	1244	14	440	1.2	14	623	15.4	45.2
Image [54]	Scene	1200	800	5	294	1.2	249	345	1.2	3.1
Medical [55]	Text	645	333	45/34	1449	1.2	0	170	60.9	230.2
PlantPseAAC [53]	Biology	588	390	12	440	1.1	12	172	6.7	21.8
Scene [56]	Scene	1211	1196	6	294	1.1	165	277	1.3	4.8
Stackex_coffee [5]	Text	151	74	123/63	1763	2.0	0	32	22.6	105.6
TMC2007-500 [57]	Text	21519	7077	22	500	2.2	304	12876	17.1	27.6
Yeast [34]	Biology	1500	917	14	103	4.2	21	1128	7.3	9.0
Yelp [58]	Text	6724	3281	5	671	1.8	580	4263	2.8	3.7

size of the class (i.e., this value is 1 for the largest class and larger for other classes). MeanIR is the mean over all the classes. Mean class imbalance ratio ("MeanCIR") denotes mean imbalance as in [60], where the imbalance of a class is computed by dividing the number of negative samples by the number of positive samples if the number of negative samples is larger and vice versa if the number of positive samples is larger. MeanCIR is the mean over all the classes. Thus, meanIR measures the imbalance between classes, which is our main interest. MeanCIR, on the other hand, focuses on the imbalance between positive and negative samples and maybe high even if all the classes have equal size.

We centralized the datasets and, for non-LDA-based techniques, we centralized also the label matrix used for training. We deleted some instances without labels or with NaN values. Some of the datasets have empty classes with no samples in either train or test set. For such datasets, we used all the samples and the full label matrix for training, but for computing the evaluation metrics we considered only classes with at least one test sample. If the number of test classes for a dataset is lower than the overall class number, we show also the number of test classes in the "Classes" column of Table I.

B. Evaluation Metrics

We adopt seven different evaluation metrics [61] to evaluate the performance of our proposed algorithm. Here, we denote the ground-truth label matrix for the M test samples as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$, where the ith column $\mathbf{y}_i \in \mathbb{R}^C$ represents the label vector of test sample $\mathbf{x}_i$. The multilabel classifiers give as their outputs for each input vector $\mathbf{x}_i$, a vector $\hat{\mathbf{p}}_i = f(\mathbf{x}_i)$, where $\hat{p}_{i,c}$ denotes the membership of instance i in class c. This is then converted to a binary predicted label vector $\hat{\mathbf{y}}_i$ by thresholding. $\mathcal{L}_i = \{\text{sort}_c(\hat{\mathbf{p}}_i)\}$ denotes an ordered list of classes ranked in the order of descending probability in $\hat{\mathbf{p}}_i$. $\mathcal{I}(\mathbf{y}_i)$ is used to denote the indices of relevant classes in $\mathbf{y}_i$ and $\neg\mathcal{I}(\mathbf{y}_i)$ denotes the indices of negative classes in $\mathbf{y}_i$. We use ($\downarrow$) to denote metrics, where lower values indicate better results and ($\uparrow$) in the opposite case.

1) *Ranking loss ($\downarrow$)* evaluates for each item i relevant versus irrelevant class pair and gives the fraction of pairs, where the irrelevant class if ranked above the relevant one. Here, we use m to denote the number of relevant classes in $\mathbf{y}_i$ and $n = C - m$

$$\text{ranking\_loss}_i = \frac{|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|}{m * n} \tag{41}$$

$$\text{ranking\_loss} = \frac{\sum_{i=1}^{M} \text{ranking\_loss}_i}{M} \tag{42}$$

where $|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|$ is used to denote the count of wrong rankings for item i.

2) *One error ($\downarrow$)* shows how often the top-ranked class for an item is not among the positive ground-truth labels

$$\text{one\_error}_i = \begin{cases} 0, & \text{if } \mathcal{L}_i[1] \in \mathcal{I}(\mathbf{y}_i) \\ 1, & \text{otherwise} \end{cases} \tag{43}$$

where $\mathcal{L}_i[1]$ denotes the first class in the sorted list $\mathcal{L}_i$

$$\text{one\_error} = \frac{\sum_{i=1}^{M} \text{one\_error}_i}{M}. \tag{44}$$

3) *Normalized coverage ($\downarrow$)* demonstrates how far on average in the predicted label ranking $\mathcal{L}_i$ one needs to go to cover all the ground-truth labels of an instance

$$\text{coverage} = \frac{\sum_{i=1}^{M} \max_j \{j | \mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\} - 1}{M * (C - 1)} \tag{45}$$

where $\{j|\mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\}$ gives the positions of relevant classes $\mathcal{I}(\mathbf{y}_i)$ in the ordered list $\mathcal{L}$.

4) *Macro-AUC ($\uparrow$)* is the average area under ROC curves (AUC) for different classes [61]. The ROC curve uses the true-positive rate and false-positive rate, which may be unreliable in the cases, where very rare classes are present (high meanCIR) [62].

5) *Micro-AUC ($\uparrow$)* is the area under ROC curves (AUC) averaged over the full predicted label matrix $\hat{\mathbf{Y}}$ [61].

6) *Macro-F1 ($\uparrow$)* shows the average $F1$ value on each class

$$\text{macro}F1 = \frac{2}{C} \sum_{c=1}^{C} \frac{\text{precision}_c * \text{recall}_c}{\text{precision}_c + \text{recall}_c} \tag{46}$$

where $\text{precision}_c = \text{TP}_c/(\text{TP}_c + \text{FP}_c)$ and $\text{recall}_c = \text{TP}_c/(\text{TP}_c + \text{FN}_c)$ are precision and recall for class c,

TABLE II
SUMMARY OF THE EVALUATION METRIC PROPERTIES

Metric	Optimized by label-wise eff.	Optimized by instance-wise eff.	Uses threshold
Ranking loss	✓		
One error	✓		
Normalized coverage	✓		
Macro-AUC		✓	
Micro-AUC	✓		
Macro-FI		✓	✓
Micro-FI			✓

and TP_c, FP_c, and FN_c are the number of true positives, false positives, and false negatives for class c.

7) *Micro-F1 (↑)* indicates the overall $F1$ score averaged over the full predicted label matrix $\hat{\mathbf{Y}}$

$$\mathrm{micro}F1 = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{47}$$

where precision = TP/(TP+FP) and recall = TP/(TP+FN) and TP, FP, and FN are the number of true positives, false positives, and false negatives predictions in the predicted label matrix $\hat{\mathbf{Y}}$.

Some characteristics of the used metrics are summarized in Table II following the analysis provided in [61]. Most of the metrics are based on the predicted membership vectors $\hat{\mathbf{p}}_i$, while the last two use the predicted class labels that can be obtained from the predicted memberships by setting a threshold. It is possible to get different predicted labels from the same $\hat{\mathbf{p}}_i$ with different thresholds, but this does not depend on the input features, that is, the quality of the dimensionality reduction techniques. Therefore, the metrics based on the predicted memberships are well suited for evaluating the differences of the dimensionality reduction techniques.

Most of the metrics can be optimized by *labelwise effective* classifiers, which roughly means that the classifier can give higher membership values for the relevant classes than for the irrelevant classes for every sample. *Instancewise effective* classifiers, on the other hand, can distinguish between relevant and irrelevant samples for each class. Some classifiers, such as Micro-FI, are optimized only by *double effective* classifiers that are both labelwise and instancewise effective. The metrics that optimize instancewise effectiveness give more weight to the samples in smaller classes and, thus, are suitable for evaluating the performance in imbalanced (high meanIR) datasets, when it is not desired to obtain an overall high performance by predicting the majority classes correctly and failing in the rare classes. Due to the aforementioned unreliability of ROC curves in the presence of a very small class (high meanCIR), we use macro-F1 as the main metric for imbalance-aware evaluation.

C. Experimental Setup

We carried out all the experiments using two multilabel classifiers applied to the projected data: 1) multilabel k-nearest neighbor classifier (ML-kNN) [54] and 2) multioutput linear ridge regressor (LRR) [2], [63]. ML-kNN utilizes the k-nearest neighbor algorithm and maximum a posterior (MAP) principle to tackle the multilabel categorization task. ML-kNN first estimates prior and posterior probabilities of each instance i for each class c from a training dataset based on frequency counting [54]. Then, the predicted probabilities on a test dataset are calculated using the Bayesian rule. In our work, the predicted labels were obtained by setting a threshold (≥ 0.5) for the predicted probabilities.The hyperparameter k of ML-kNN was set to 15 as in [14]. As multilabel classification is a specific case of multitarget regression [64], the multioutput LRR can be trained to solve the multilabel classification tasks. In our work, we used the LRR classifier with a hyperparameter $\mu = 0.1$. The predicted labels were obtained by setting a threshold (≥ 0) for the predicted values from the LRR classifiers.

For comparisons, we used the following LDA-based dimensionality reduction techniques: DMLDA [21], wMLDAc, wMLDAb, wMLDAe, wMLDAf, and wMLDAd [14], where the subscripts denote the types of prior information used as weight factors following Section III-A1. Note that wMLDAc is equivalent to the original MLDA [12]. For all the LDA-based methods, we solved the regularized generalized eigenproblem (39) with $\epsilon = 0.1$. After solving the eigenproblem, we kept the eigenvectors corresponding to the top 0.999 informative eigenvalues to form the projection matrix $\mathbf{W}$. Besides the LDA-based methods, we conducted experiments with five other dimensionality reduction techniques: PCA, CCA [16], MLSI [17], MDDM_d [18], and MVMD [19]. We used the MATLAB codes provided for [14][1] in the comparative experiments and exploit the relevant parts also in the implementation of our proposed method.

D. Classification Results and Analysis

1) Comparisons of Different Variants of SMLDA and wMLDA: We first compare the different variants of our proposed SMLDA approach. Furthermore, we compare our methods against the variants of wMLDA that use the same prior information types directly as weights. We show the results using the ranking loss evaluation metric in Tables III and IV of the main paper and the results using the six other evaluation metrics in Tables I–XII of the supplementary material. In each table, we place next to each other the variants of SMLDA and wMLDA with the same prior information type and highlight the better approach for each dataset. The prior information for SMLDAm was proposed by us and has not been previously used with wMLDA. Therefore, we do not show such a comparison for it.

We first observe that our proposed SMLDA variants clearly outperform the corresponding wMLDA variants. In all test cases by both classifiers and any evaluation metric, the average performance of the proposed approach is better. This clearly confirms the value of using the probabilistic saliency estimation instead of just using the same prior information type directly as a weight as in wMLDA.

Next, we observe that there are no major differences among the variants of SMLDA. Therefore, we do not recommend using SMLDAd or SMLDAf because the fuzzy and dependence-based prior information types are computationally much more expensive than the other prior information types. Among the remaining variants, we select SMLDAc as our default variant.

TABLE III
COMPARISON OF DIFFERENT VARIANTS OF THE PROPOSED METHOD RESULTS WITH ML-KNN USING RANKING LOSS (↓)

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.164	**0.149**	**0.151**	0.152	0.153	**0.149**	0.151	**0.150**	0.147	**0.146**	0.147
Birds	0.217	**0.200**	0.206	**0.197**	**0.193**	0.197	0.201	**0.196**	0.232	**0.204**	0.204
CHD_49	0.212	**0.195**	**0.200**	0.206	0.198	**0.194**	**0.195**	0.200	0.226	**0.206**	0.205
Cal500	0.190	**0.187**	**0.187**	0.187	0.188	**0.187**	0.187	**0.187**	**0.186**	0.188	0.186
Corel16k(001)	0.190	**0.187**	**0.186**	0.186	**0.187**	0.187	**0.187**	0.187	0.186	**0.184**	0.182
Emotions	**0.173**	0.190	**0.162**	0.187	**0.164**	0.177	0.182	**0.177**	0.205	**0.184**	0.182
Enron	0.218	**0.142**	0.188	**0.145**	0.177	**0.142**	0.178	**0.142**	0.161	**0.139**	0.142
Eukaryote	0.122	**0.121**	0.122	**0.121**	**0.121**	0.121	**0.120**	0.121	**0.119**	0.121	0.120
Human	0.173	**0.160**	0.172	**0.162**	0.171	**0.162**	0.172	**0.159**	0.171	**0.162**	0.157
Image	0.193	**0.173**	0.199	**0.160**	0.195	**0.167**	0.199	**0.166**	0.203	**0.162**	0.172
Medical	0.071	**0.060**	0.066	**0.059**	0.065	**0.060**	0.064	**0.059**	0.071	**0.058**	0.057
PlantPseAAC	0.280	**0.228**	0.260	**0.230**	0.284	**0.225**	0.291	**0.229**	0.271	**0.234**	0.224
Scene	0.135	**0.088**	0.137	**0.087**	0.135	**0.089**	0.135	**0.088**	0.132	**0.089**	0.092
Stackex_coffee	**0.241**	0.273	**0.268**	0.272	**0.269**	0.272	**0.271**	0.272	0.284	**0.270**	0.275
TMC2007	0.027	**0.026**	0.026	**0.026**	0.026	**0.026**	0.026	**0.026**	0.028	**0.026**	0.027
Yeast	0.183	**0.178**	0.185	**0.177**	0.183	**0.178**	0.185	**0.178**	0.185	**0.177**	0.178
Yelp	0.126	**0.124**	0.130	**0.123**	0.126	**0.125**	**0.125**	0.126	0.139	**0.131**	0.139
Average	0.171	**0.158**	0.167	**0.158**	0.167	**0.156**	0.169	**0.157**	0.173	**0.158**	0.158
	Statistical analysis: Friedman: $p =$ **4.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	25.0	X	**25.0**	-66.0	36.0	-53.0	**30.0**	-57.0	**10.0**	-73.0	-67.0

TABLE IV
COMPARISON OF DIFFERENT VARIANTS OF THE PROPOSED METHOD RESULTS WITH LRR USING RANKING LOSS (↓)

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.120	**0.115**	0.120	**0.115**	0.120	**0.115**	0.119	**0.115**	0.124	**0.114**	0.112
Birds	0.332	**0.268**	0.321	**0.265**	0.321	**0.270**	0.318	**0.270**	0.321	**0.263**	0.258
CHD_49	0.209	**0.207**	0.208	**0.203**	0.208	**0.204**	0.208	**0.204**	0.213	**0.196**	0.189
Cal500	**0.242**	0.267	**0.250**	0.267	**0.266**	0.266	**0.265**	0.267	**0.194**	0.267	0.266
Corel16k(001)	**0.201**	0.202	0.206	**0.202**	0.204	**0.202**	0.204	**0.202**	**0.190**	0.199	0.197
Emotions	0.172	**0.163**	0.166	**0.161**	0.170	**0.168**	**0.168**	0.168	0.171	**0.162**	0.159
Enron	0.360	**0.198**	0.344	**0.195**	0.327	**0.196**	0.326	**0.196**	0.306	**0.195**	0.189
Eukaryote	0.129	**0.125**	0.130	**0.126**	0.129	**0.125**	0.129	**0.125**	0.128	**0.125**	0.123
Human	0.199	**0.179**	0.203	**0.181**	0.200	**0.178**	0.199	**0.178**	0.194	**0.174**	0.171
Image	0.208	**0.174**	0.203	**0.174**	0.205	**0.172**	0.203	**0.172**	0.203	**0.175**	0.188
Medical	0.044	**0.027**	0.036	**0.027**	0.035	**0.027**	0.033	**0.027**	0.044	**0.026**	0.028
PlantPseAAC	0.356	**0.339**	0.352	**0.336**	0.352	**0.339**	0.351	**0.339**	0.352	**0.338**	0.329
Scene	0.137	**0.091**	0.136	**0.092**	0.137	**0.092**	0.136	**0.091**	0.133	**0.092**	0.092
Stackex_coffee	0.199	**0.157**	**0.158**	0.160	**0.159**	0.160	0.188	**0.162**	0.188	**0.156**	0.157
TMC2007	0.040	**0.040**	**0.038**	0.040	**0.039**	0.040	**0.039**	0.040	0.047	**0.040**	0.042
Yeast	0.184	**0.178**	0.182	**0.178**	0.184	**0.178**	0.185	**0.178**	0.188	**0.178**	0.177
Yelp	0.137	**0.136**	0.137	**0.136**	0.137	**0.135**	0.137	**0.135**	0.148	**0.142**	0.147
Average	0.192	**0.169**	0.188	**0.168**	0.188	**0.169**	0.189	**0.169**	0.185	**0.167**	0.166
	Statistical analysis: Friedman: $p =$ **1.8e-10**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	14.0	X	**16.0**	-53.0	**5.0**	-69.5	**3.0**	-56.0	**23.0**	**-31.0**	-44.0

2) Comparisons Against Competing Dimensionality Reduction Techniques: We then compare wMLDAc, which we recommend using as our default variant, against other competing dimensionality reduction techniques. Here, we consider five non-LDA-based techniques: 1) PCA; 2) CCA; 3) MLSI; 4) MDDMp; and 5) MVMD, along with DMLDA, MLDA, which is equivalent to wMLDAc and uses the same prior information as our proposed variant wMLDAc, and wMLDAd, which was the proposed wLMDA variant in [14]. We provide the results in Tables V and VI of the main paper and Tables XIII–XXIV of the supplementary material.

The results show that our proposed method has the best average performance with ML-kNN evaluated by all the performance metrics and with LRR evaluated by macro-F1. MDDMp is the best performing competing method. However, in all cases, our proposed approach achieves a similar performance, while our method is clearly better when evaluated with macro-F1. Our proposed method also clearly outperforms other LDA-based techniques.

We then focus on the most imbalanced datasets evaluated by our main metric for imbalanced classification, macro-F1. We collect from Tables XVII and XVIII of the supplementary material the results for the classes having meanIR over 15 and provide them in Tables VII and VIII. Our proposed method has the best average performance with ML-kNN and the second best with LRR, which shows that the proposed method indeed can help to deal with class imbalance.

3) Statistical Analysis of the Results: To evaluate whether the observed differences are statistically significant, we followed the recommendations of [65]. We first applied to each table the Friedman test, which is a rank-based nonparametric test showing whether the differences are overall significant. At the bottom of each table, we report the Friedman p value. We have highlighted the value if it shows that the null hypothesis can be rejected at the 0.05 significance level. Next, we perform the Wilcoxon sign-ranks test to evaluate the pairwise differences between the methods. This test ranks the differences between two classifiers ignoring the signs and uses the ranks to determine value T as described, for example, in [65]. Finally, the T value is compared to a critical value that depends on the number of datasets. In our experiments, we used 17 datasets, which means that the null hypothesis

TABLE V
COMPARATIVE RESULTS WITH ML-KNN USING RANKING LOSS (↓)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.204	0.197	0.199	**0.116**	0.186	0.271	0.164	0.147	0.149
Birds	0.323	0.203	0.323	0.322	0.322	0.248	0.217	0.232	**0.200**
CHD_49	0.224	0.212	0.214	0.209	0.225	0.224	0.212	0.226	**0.195**
Cal500	0.183	0.187	0.184	**0.182**	0.183	0.187	0.190	0.186	0.187
Corel16k(001)	0.198	0.188	0.196	**0.185**	0.198	0.197	0.190	0.186	0.187
Emotions	0.299	0.178	0.299	0.301	0.295	0.245	**0.173**	0.205	0.190
Enron	0.133	0.170	0.135	**0.124**	0.136	0.191	0.218	0.161	0.142
Eukaryote	0.113	0.126	0.113	**0.106**	0.111	0.141	0.122	0.119	0.121
Human	0.159	0.178	0.159	**0.149**	0.158	0.191	0.173	0.171	0.160
Image	0.167	0.201	0.170	0.186	**0.166**	0.284	0.193	0.203	0.173
Medical	0.057	0.076	**0.039**	0.051	0.058	0.072	0.071	0.071	0.060
PlantPseAAC	0.197	0.277	0.198	**0.180**	0.198	0.258	0.280	0.271	0.228
Scene	0.084	0.141	0.083	0.102	**0.077**	0.234	0.135	0.132	0.088
Stackex_coffee	0.279	0.304	0.259	0.257	0.276	0.298	**0.241**	0.284	0.273
TMC2007	0.035	0.026	0.035	0.030	0.030	0.038	0.027	0.028	**0.026**
Yeast	0.174	0.184	**0.173**	0.179	0.174	0.188	0.183	0.185	0.178
Yelp	0.178	**0.117**	0.171	0.148	0.176	0.139	0.126	0.139	0.124
Average	0.177	0.174	0.174	0.166	0.175	0.200	0.171	0.173	**0.158**
	Statistical analysis: Friedman: $p =$ **1.8e-04**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	52.0	**16.0**	63.0	-74.0	61.0	**0.0**	**25.0**	**10.0**	X

TABLE VI
COMPARATIVE RESULTS WITH LRR USING RANKING LOSS (↓)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.117	0.120	0.117	**0.079**	0.091	0.120	0.120	0.124	0.115
Birds	0.236	0.301	0.288	**0.171**	0.199	0.318	0.332	0.321	0.268
CHD_49	0.208	0.210	0.208	**0.196**	0.205	0.210	0.209	0.213	0.207
Cal500	0.258	0.269	0.265	0.248	0.250	0.245	0.242	**0.194**	0.267
Corel16k(001)	0.208	0.208	0.208	0.195	0.208	0.206	0.201	**0.190**	0.202
Emotions	0.163	0.167	0.163	0.174	0.177	0.166	0.172	0.171	**0.163**
Enron	0.250	0.324	0.332	**0.121**	0.138	0.400	0.360	0.306	0.198
Eukaryote	0.134	0.130	0.134	**0.111**	0.120	0.131	0.129	0.128	0.125
Human	0.211	0.209	0.210	**0.157**	0.185	0.210	0.199	0.194	0.179
Image	0.206	0.207	0.217	0.198	0.177	0.223	0.208	0.203	**0.174**
Medical	0.031	0.039	0.057	0.025	**0.024**	0.063	0.044	0.044	0.027
PlantPseAAC	0.340	0.351	0.343	**0.194**	0.315	0.362	0.356	0.352	0.339
Scene	0.136	0.136	0.138	0.097	**0.088**	0.141	0.137	0.133	0.091
Stackex_coffee	0.170	0.168	0.169	**0.157**	0.171	0.163	0.199	0.188	0.157
TMC2007	0.038	0.037	0.038	0.049	0.048	**0.037**	0.040	0.047	0.040
Yeast	0.184	0.183	0.184	0.180	0.179	0.182	0.184	0.188	**0.178**
Yelp	0.130	0.129	0.130	0.165	0.135	**0.129**	0.137	0.148	0.136
Average	0.178	0.188	0.188	**0.148**	0.159	0.195	0.192	0.185	0.169
	Statistical analysis: Friedman: $p =$ **6.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	37.0	**11.0**	**16.0**	-44.0	-55.0	**20.0**	**14.0**	**23.0**	X

TABLE VII
COMPARATIVE RESULTS WITH ML-KNN USING MACRO-F1 (↑)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.056	0.050	0.055	**0.062**	0.055	0.051	0.051	0.056	0.051
Corel16k(001)	0.013	0.030	0.017	0.036	0.018	0.018	**0.037**	0.034	0.036
Enron	0.046	0.073	0.042	**0.095**	0.054	0.012	0.039	0.036	0.062
Eukaryote	0.053	0.074	0.053	0.060	0.065	0.002	0.090	**0.092**	0.072
Human	0.043	0.146	0.041	0.095	0.071	0.001	0.145	0.133	**0.159**
Medical	0.219	0.294	**0.307**	0.280	0.226	0.186	0.259	0.263	0.302
Stackex_coffee	0.000	0.023	0.017	0.013	0.000	0.010	0.036	0.040	**0.048**
Average	0.061	0.099	0.076	0.092	0.070	0.040	0.094	0.093	**0.104**
	Statistical analysis: Friedman: $p =$ **2.7e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	1.0	7.0	3.0	7.0	**1.0**	**0.0**	8.0	6.0	X

can be rejected at 0.01 significance level if $T_1 \leq 23$ and at 0.05 significance level if $T_2 \leq 34$. For seven datasets, as in Tables VII and VIII, $T_2 \leq 2$. We applied the Wilcoxon sign-ranks test between our proposed SMLDAc method and every other method dimensionality reduction technique. We give these values at the bottom of every table and bold the values if they show that the difference between the methods is statistically significant at a 0.05 significance level. Negative

TABLE VIII
COMPARATIVE RESULTS WITH LRR USING MACRO-F1 (↑)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.122	0.125	0.126	0.120	0.120	0.104	0.103	0.070	**0.127**
Corel16k(001)	0.043	0.044	0.043	0.035	0.041	**0.044**	0.042	0.037	0.044
Enron	**0.123**	0.117	0.121	0.086	0.101	0.097	0.101	0.095	0.113
Eukaryote	0.113	**0.119**	0.113	0.097	0.111	0.117	0.119	0.117	0.111
Human	0.143	0.149	0.145	0.136	0.151	0.148	0.147	0.150	**0.156**
Medical	0.531	**0.551**	0.489	0.444	0.487	0.488	0.440	0.443	0.536
Stackex_coffee	0.171	0.179	0.186	0.124	0.165	0.190	0.144	0.159	**0.196**
Average	0.178	**0.184**	0.175	0.149	0.168	0.170	0.156	0.153	0.183
	Statistical analysis: Friedman: $p =$ **1.2e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	7.0	-14.0	7.0	**0.0**	**1.0**	4.0	**2.0**	**1.0**	X

TABLE IX
SUMMARY OF THE WILCOXON SIGNED-RANKS TEST RESULTS: THE NUMBER OF TIMES WHEN SMLDAC WAS BETTER IN A STATISTICALLY SIGNIFICANT WAY

	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAb	wMLDAe	wMLDAf	wMLDAd
ML-kNN	4/7	5/7	2/7	1/7	2/7	7/7	4/7	4/7	1/7	4/7	7/7
LRR	1/7	4/7	5/7	1/7	0/7	4/7	7/7	5/7	7/7	5/7	7/7
Total	5/14	9/14	7/14	2/14	2/14	11/14	11/14	9/14	8/14	9/14	14/14

values indicate that the other method was performing better than SMLDAc.

The results of the Friedman test show that the overall differences are statistically significant in most cases. The only exceptions among 28 result tables are Tables I, II, XI, and XIV in the supplementary material. Tables I and II in the supplementary material, compare the variants of the proposed method using ML-kNN and LRR with one error evaluation metric. Table XI in the supplementary material, compares the variants of the proposed method using ML-kNN with the Micro-F1 evaluation metric. Table XIV in the supplementary material, compares SMLDAc against competing methods using LRR and one error evaluation metric.

The results of the Wilcoxon signed-ranks test confirm the good performance of our proposed SMLDAc. There is no such case, where a competing method would outperform SMLDAc in a statistically significant manner (only another variant of our proposed method, SMLDAd, can do this in two cases). On the other hand, SMLDAc can outperform every competing method in a statistically significant manner at least twice. The results of the conducted Wilcoxon signed-ranks test are summarized in Table IX showing the number of times when a statistically significant difference was detected between SMLDAc and all competing methods.

V. CONCLUSION

In this article, we proposed a novel probabilistic framework for the LDA-related dimensionality reduction algorithm aiming to improve the performance of multilabel classifiers on various multilabel datasets. The probabilistic approach uses an affinity matrix to ensure similar results for similar instances and a prior information matrix to integrate prior information on the prominence of each instance for each class. Our solution can alleviate the data imbalance problem, which is commonly encountered in multilabel datasets, as the weight factor vectors are calculated separately for each class. Our method can also alleviate the common overcounting problem. We proposed variants of our methods using different prior information matrices based on both labels and features.

We used seven metrics to evaluate the performance of our method with competing methods on 17 multilabel datasets. The experimental results showed that our method enhanced the classification performance compared to the competing algorithms and handles imbalanced classification well. Our algorithm is still based on the linear subspace learning technique. In the future, we will make a nonlinear extension using the kernel trick.

REFERENCES

[1] L. Li, H. Wang, X. Sun, B. Chang, S. Zhao, and L. Sha, "Multi-label text categorization with joint learning predictions-as-features method," in *Proc. Conf. Empirical Methods Nat. Language Process.*, 2015, pp. 835–839.

[2] C. Tan, S. Chen, G. Ji, and X. Geng, "Multilabel distribution learning based on multioutput regression and manifold learning," *IEEE Trans. Cybern.*, early access, Oct. 23, 2020, doi: 10.1109/TCYB.2020.3026576.

[3] J. Read, L. Martino, and J. Hollmén, "Multi-label methods for prediction with sequential data," *Pattern Recognit.*, vol. 63, pp. 45–55, Sep. 2016.

[4] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music by emotion," *EURASIP J. Audio Speech Music Process.*, vol. 2011, no. 1, pp. 1–9, 2011. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/1687-4722-2011-426793

[5] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, Sep. 2015.

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

[7] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification*. Cham, Switzerland: Springer, 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-41111-8_1#citeas

[8] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 4, no. 6, pp. 411–444, 2014.

[9] W. Siblini, P. Kuntz, and F. Meyer, "A review on dimensionality reduction for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 8, pp. 839–857, Mar. 2021.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Regularized extreme learning machine for multi-view semi-supervised action recognition," *Neurocomputing*, vol. 145, pp. 250–262, Dec. 2014.

[11] H. Wang, L. Yan, H. Huang, and C. Ding, "From protein sequence to protein function via multi-label linear discriminant analysis," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 14, no. 3, pp. 503–513, May/Jun. 2017.

[12] H. Wang, C. Ding, and H. Huang, *Multi-Label Linear Discriminant Analysis* (LNCS 6316). Heidelberg, Germany: Springer, 2010, pp. 126–139. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-15567-3_10#citeas

[13] L. Xu, A. Iosifidis, and M. Gabbouj, "Weighted linear discriminant analysis based on class saliency information," in *Proc. Int. Conf. Image Process. (ICIP)*, 2018, pp. 2306–2310.

[14] J. Xu, "A weighted linear discriminant analysis framework for multi-label feature extraction," *Neurocomputing*, vol. 275, pp. 107–120, Jan. 2018.

[15] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philosop. Trans. Roy. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016.

[16] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.

[17] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2005, pp. 258–265.

[18] Y. Zhang and Z. H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Trans. Knowl. Disc. Data*, vol. 4, no. 3, pp. 1–21, 2010.

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

[20] C. H. Park and M. Lee, "On applying linear discriminant analysis for multi-labeled problems," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 878–887, 2008.

[21] M. Oikonomou and A. Tefas, "Direct multi-label linear discriminant analysis," *Commun. Comput. Inf. Sci.*, vol. 383, no. 1, pp. 414–423, 2013.

[22] Y. Yuan, K. Zhao, and H. Lu, "Multi-label linear discriminant analysis with locality consistency," in *Proc. Int. Conf. Neural Inf. Process.*, 2014, pp. 386–394.

[23] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.

[24] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[25] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–9.

[26] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2100–2110, Oct. 2017.

[27] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognit.*, vol. 37, no. 5, pp. 857–874, 2004.

[28] E. K. Tang, P. N. Suganthan, X. Yao, and A. K. Qin, "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognit.*, vol. 38, no. 4, pp. 485–493, 2005.

[29] E. Tang, P. Suganthan, and X. Yao, "Generalized LDA using relevance weighting and evolution strategy," in *Proc. Congr. Evol. Comput.*, vol. 1, 2005, pp. 2230–2234.

[30] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, Apr. 2009.

[31] C. Aytekin, A. Iosifidis, and M. Gabbouj, "Probabilistic saliency estimation," *Pattern Recognit.*, vol. 74, pp. 359–372, Feb. 2018.

[32] H. Ahmed, J. Mohamed, and Z. Noureddine, "Face recognition systems using relevance weighted two dimensional linear discriminant analysis algorithm," *J. Signal Inf. Process.*, vol. 3, no. 1, pp. 130–135, 2012.

[33] Q. Wu, M. Tan, H. Song, J. Chen, and M. K. Ng, "ML-FOREST: A multi-label tree ensemble method for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2665–2680, Oct. 2016.

[34] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, vol. 14, no. 4, pp. 897–911, 1992.

[35] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, 2007, pp. 451–456.

[36] X. Lin and X.-W. Chen, "Mr.KNN—Soft relevance for multi-label classification." in *Proc. 19th ACM Conf. Inf. Knowl. Manage.*, 2010, pp. 349–358.

[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[38] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[39] C. Li *et al.*, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[40] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli, "A feature-based approach for saliency estimation of omni-directional images," *Signal Process. Image Commun.*, vol. 69, pp. 53–59, Mar. 2018.

[41] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 37, 2011, pp. 409–416.

[42] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[43] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.

[44] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, Jun. 2019.

[45] A. Gretton, O. Bousquet, A. Smola, and B. Schlkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," *Proc. 16th Int. Conf. Algorithmic Learn. Theory (ALT)*, 2005, pp. 63–77.

[46] I. Katakis, G. Tsoumakas, and V. Ioannis, "Multilabel text classification for automated tag suggestion," in *Proc. ECML/PKDD Disc. Challenge*, 2008, pp. 1107–1135.

[47] F. Briggs *et al.*, "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2013, pp. 1–8.

[48] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[49] H. Shao, G. Li, G. Liu, and Y. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine," *Sci. China Inf. Sci.*, vol. 56, no. 5, pp. 1–13, 2013.

[50] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, no. 6, pp. 1107–1135, 2003.

[51] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD Workshop Min. Multidimensional Data (MMD)*, 2008, pp. 30–44. [Online]. Available: http://lpis.csd.auth.gr/publications/tsoumakas-mmd08.pdf

[52] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases (ECML PKDD)*, vol. 5782, 2009, pp. 254–269. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-04174-7_17#citeas

[53] J. Xu, "Fast multi-label core vector machine," *Pattern Recognit.*, vol. 46, no. 3, pp. 885–898, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312003950

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

[55] J. P. Pestian *et al.*, "A shared task involving multi-label classification of clinical free text," in *Proc. ACL Workshop BioNLP Biol. Transl. Clin. Lang. Process.*, Jun. 2007, pp. 97–104.

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

[57] A. N. Srivastava and B. Zane-Ulman, "Discovering recurring anomalies in text reports regarding complex space systems," in *Proc. IEEE Aerosp. Conf.*, 2005, pp. 3853–3862.

[58] H. Sajnani, V. Saini, K. Kumar, E. Gabrielova, P. Choudary, and C. Lopes. (2013). *Classifying Yelp Reviews Into Relevant Categories.* [Online]. Available: http://www.ics.uci.edu/ vpsaini/.

[59] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "A first approach to deal with imbalance in multi-label datasets," in *Hybrid Artificial Intelligent Systems*, J.-S. Pan, M. M. Polycarpou, M. Woźniak, A. C. P. L. F. de Carvalho, H. Quintián, and E. Corchado, Eds.

Heidelberg, Germany: Springer, 2013, pp. 150–160. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-40846-5_16#citeas
[60] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu, "Towards class-imbalance aware multi-label learning," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*, 2015, pp. 4041–4047.
[61] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3780–3788.
[62] J. Davis and M. Goadrich, "The relationship between precision-recall and RoC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, New York, NY, USA, 2006, pp. 233–240. [Online]. Available: https://doi.org/10.1145/1143844.1143874
[63] H. Borchani, G. Varando, C. Bielza, and B. Monte, "A survey on multi-output regression," *Interdiscipl. Rev. Data Min. Knowl. Discov.*, vol. 5, no. 5, pp. 216–233, 2015.
[64] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: Treating targets as inputs," *Mach. Learn.*, vol. 104, no. 1, pp. 55–98, 2016.
[65] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.

Lei Xu (Student Member, IEEE) received the B.S.E.E. degree from East China Normal University, Shanghai, China, in 2006, and the M.S.E.E. degree from the Tampere University of Technology, Tampere, Finland, in 2017. She is currently pursuing the Ph.D. degree with Tampere University, Tampere.

From 2006 to 2013, she was an Engineer in Shanghai, where she was involved with on-train communication systems design. Her current research interests include artificial intelligence, data science, and machine learning.

Jenni Raitoharju (Member, IEEE) received the Ph.D. degree from the Tampere University of Technology, Tampere, Finland, in 2017.

She works as a Senior Research Scientist with the Finnish Environment Institute, Jyväskylä, Finland. She has coauthored 25 international journal papers and 34 papers in international conferences. She currently leads two research projects funded by the Academy of Finland, focusing on automatic taxa identification. Her research interests include machine learning and pattern recognition methods along with applications in biomonitoring and autonomous systems.

Dr. Raitoharju is the Chair of Young Academy Finland from 2019 to 2021.

Alexandros Iosifidis (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Democritus University of Thrace, Komotini, Greece, in 2008 and 2010, respectively, and the Ph.D. degree from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2014.

He is currently an Associate Professor with Aarhus University, Aarhus, Denmark. He has coauthored 84 articles in international journals and 100 papers in international conferences and workshops. His research interests include topics of neural networks and statistical machine learning finding applications in computer vision, financial engineering, and graph analysis problems.

Prof. Iosifidis's work has received the H.C. Oersted Young Researcher Prize 2018 and the EURASIP Early Career Award 2021. He served as an Officer of the Finnish IEEE SP/CAS Chapter from 2016 to 2018. He is currently a member of the EURASIP Technical Area Committee on Visual Information Processing. He serves as the Associate Editor-in-Chief for *Neurocomputing*, as an Area Editor for *Signal Processing: Image Communications*, and an Associate Editor for *BMC Bioinformatics*.

Moncef Gabbouj (Fellow, IEEE) received the B.S. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1986 and 1989, respectively.

He is a Professor of Signal Processing with the Department of Computing Sciences, Tampere University, Tampere, Finland. He was an Academy of Finland Professor from 2011 to 2015. He is the Finland Site Director of the NSF IUCRC funded Center for Visual and Decision Informatics and leads the Artificial Intelligence Research Task Force of the Ministry of Economic Affairs and Employment funded Research Alliance on Autonomous Systems. His research interests include big data analytics, multimedia content-based analysis, indexing and retrieval, artificial intelligence, machine learning, pattern recognition, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Prof. Gabbouj is the past Chairman of the IEEE CAS TC on DSP and a committee member of the IEEE Fourier Award for Signal Processing. He has served as an associate editor and a guest editor for many IEEE and international journals and a Distinguished Lecturer for the IEEE CASS. He is a member of the Academia Europaea and the Finnish Academy of Science and Letters.