applied sciences

MDPI

Article

An Effective Multi-Label Feature Selection Model Towards Eliminating Noisy Features

Jun Wang [1], Yuanyuan Xu [2], Hengpeng Xu [3], Zhe Sun [4], Zhenglu Yang [2] and Jinmao Wei [2,*]

1 College of Mathematics and Statistics Science, Ludong University, Yantai 264025, China; junwang@mail.nankai.edu.cn
2 College of Computer Science, Nankai University, Tianjin 300071, China; xuyuanyuan@mail.nankai.edu.cn (Y.X.); yangzl@nankai.edu.cn (Z.Y.)
3 Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China; xuhp@tjnu.edu.cn
4 RIKEN National Science Institute, Wako, Saitama 351-0198, Japan; zhe.sun.vk@riken.jp
* Correspondence: weijm@nankai.edu.cn

Received: 21 October 2020; Accepted: 12 November 2020; Published: 15 November 2020

Abstract: Feature selection has devoted a consistently great amount of effort to dimension reduction for various machine learning tasks. Existing feature selection models focus on selecting the most discriminative features for learning targets. However, this strategy is weak in handling two kinds of features, that is, the irrelevant and redundant ones, which are collectively referred to as noisy features. These features may hamper the construction of optimal low-dimensional subspaces and compromise the learning performance of downstream tasks. In this study, we propose a novel multi-label feature selection approach by embedding label correlations (dubbed ELC) to address these issues. Particularly, we extract label correlations for reliable label space structures and employ them to steer feature selection. In this way, label and feature spaces can be expected to be consistent and noisy features can be effectively eliminated. An extensive experimental evaluation on public benchmarks validated the superiority of ELC.

Keywords: feature selection; noise elimination; space consistency; label correlations

1. Introduction

For pattern recognition, feature selection is important for its effectiveness in reducing dimensionality. Feature selection methods are divided into supervised, semi-supervised, and unsupervised ones, according to whether the instances are labeled, partially labeled, or not [1–4]. For supervised cases, class labels are employed for measuring features' discriminative abilities. Many popular and efficient feature selection methods belong to this group [5–10]. Supervised methods are further categorized into three well-known models: filter, wrapper, and embedded [11]. In recent years, some hybrid methods have emerged that combine filter and wrapper processes for enhancing performance and reducing computational cost [12,13].

In another categorization view, existing feature selection approaches can also be grouped to single-label and multi-label ones, whose difference lies in the size of labels that each instance is related with [14]. In single-label FS, instances and labels hold many-to-one connections and the target separability is emphasized in this learning task. With the great potential and success of multi-label learning in many machine learning fields, such as text categorization [15], content annotation [16], and protein location prediction [17], multi-label feature selection has received considerable attention in recent years. We approach the supervised multi-label feature selection in this study.

In multi-label learning, label correlations are the key to combining the complicated relationships among instances, which are typically annotated with multiple labels [18,19]. The mainstream multi-label feature selection strategy is to extract label correlations (via statistical or information-based measurements) and employ them to help find the most remarkable features. A critical issue is, however, this strategy would be trapped by two kinds of features, that is, irrelevant and redundant ones. Irrelevant features represent those lowly discriminative ones. Features of this kind are loosely correlated with learning targets and even may provide misleading information. Compared with irrelevant features, redundant features seem more deceptive. They may exhibit excellent (or comparably superior) performances and mix with remarkable features. Nevertheless, redundant features also lowly contribute to enhancing the discriminative ability of the constructed low-dimensional subspace, because the learning information they provide is redundant with the already distilled information. In general, we regard both irrelevant and redundant features as noisy ones, which may confuse selection processes and compromise the learning performance of downstream tasks.

In this paper, we present an effective multi-label feature selection model by embedding label correlations to eliminate noisy features, named ELC. Our major strategy is to keep feature-label space consistent and explore reliable label structures to drive feature selection. Concretely, we qualitatively assess label correlations in the label space and embed them in feature selection. In this way, the label structure information can be maximally preserved in the constructed low-dimensional subspace, and eventually the consistency between feature and label spaces can be achieved. Furthermore, we devise an efficient framework base on the sparse multi-task learning to optimize ELC, which can help ELC find globally optimal solutions and efficiently converge.

The major contributions of this paper are as follows:

- We present a novel multi-label feature selection model to address the issue of noisy features. This model qualitatively measures label correlations and employs feature-label space consistency to steer feature selection.
- We devised a compact framework to optimize the proposed model. This framework resorts to the multi-task learning strategy and promises globally optimal solutions and efficient convergence.
- Comprehensive experiments on openly available benchmarks were conducted to validate the performance of the proposed model in feature selection and noise elimination.

The remaining parts of this paper are arranged as follows: related works are reviewed in Section 2; the proposed model ELC and its optimization framework are respectively introduced in Section 3 and Section 4; the experimental comparisons of ELC with several popular feature selection approaches are presented in Section 5; finally, conclusions are drawn in Section 6.

2. Related Work

Feature selection approaches are commonly specified to a certain recognition scenario, i.e., single-label learning or multi-label learning, because of the different concerns of the two recognition tasks. The issue of noisy feature elimination is firstly raised in single-label feature selection, focusing on removing irrelevant features and picking out discriminative ones. For example, the popular single-label feature selection family by preserving instance similarity [20] directly highly scores the most discriminative features under various statistical metrics, such as the Laplacian score [7,21], the Fisher score [6], the Hilbert–Schmidt independence criterion [22], and the trace ratio [23], just to name a few. In addition to the above similarity preservation approaches, some traditional distance or instance difference based ones can also be deemed as simply pursuing "target-specific features," such as ReliefF [10], SPEC [24,25], and SPFS [20]. This denotation arises from the fact that target-specific features are picked based only on whether they are strongly correlated with the learning targets. In other words, those features that have excellent discriminative abilities for targets will prevail. The aforementioned approaches have generally achieved excellent performance in eliminating

irrelevant features, while may experience difficulties in improving learning performance due to their scarce attention on removing redundant features.

Recently, some remarkable neural networks-based and fuzzy logic-based feature selection works have been presented, which have received extensive attention due to their excellent feature selection performances [26–28]. For example, Verikas and Bacauskiene [26] proposed a feedforward neural network-based approach to find the salient features and remove those yielding the least accurate classifications. Arefnezhad et al. [27] highly scored the features most related to the drowsiness level via an adaptive neuro-fuzzy inference system, which was devised by combining filter and wrapper feature selection approaches. Cateni et al. [28] selected the mostly relevant features for better binary classification by combining several filter approaches through a fuzzy inference system. Generally speaking, the above studies serve as excellent examples of picking out target-specific features, while still leaving aside the underlying negative effects of noisy features.

A salient but redundant feature provides little valuable learning information if selected. Although this issue is ignored by a majority of feature selection approaches, it gains attention from some information-based ones. Among them, the family based on mutual information is regarded as the mainstream redundancy removing approach. The classical mutual information [9] and its variants (e.g., conditional mutual information) [5,29] can effectively position the redundant features and remove them via a greedy search. Nevertheless, an inevitable problem is that the performances of these approaches heavily depend on their probability estimation accuracy. This problem is more complicated in high-dimensional space.

In terms of multi-label feature selection approaches, they can be roughly categorized into two families. The first family directly divides the multi-label learning into multiple subproblems and utilizes single-label feature evaluation metrics to tackle them [4]. For instance, ReliefF is tailed for multi-label learning by dividing its estimations of nearest misses and hits to eight subproblems [30]. In addition, some single-label feature evaluation strategies are also reformulated to the multi-label ones by enforcing on each subgroup, such as class separability and linear discriminant analysis [31,32]. A major drawback of the above subproblem division strategy is that it ignores label correlations, which encode the underlying label structures for recognition and play critical roles in multi-label learning.

On the other hand, the second family of multi-label feature selection can better fix this issue since it incorporates label correlations into model construction. A common strategy of this family is to evaluate instance-label pairs via specific label ranking metrics and select the features by minimizing loss functions [33–36]. While real-world label relations could be beyond pairwise situations, some high-order correlation approaches have been proposed to model complicated label structures. A feasible solution is to build a common space shared among various labels [16,33,37], which typically suffers from high costs and complex computation. It is noteworthy that in contrast to single-label feature selection approaches, the multi-label ones rarely have the issue of noisy feature elimination. A few approaches specific to ruling out irrelevant features are based on sparse regularization [38]. These approaches neglect the negative effects of redundant features and are not competent in completely removing noisy features.

To comprehensively address the above issues, we will introduce a novel multi-label feature selection model in Section 3, which can effectively filter both kinds of noisy features (including irrelevant and redundant ones) and select the remarkable ones. The proposed model adopts a statistical metric to measure target-related feature redundancy and dispense with any probability estimation. Furthermore, this model extracts label correlations and keeps feature-label space consistency to guide feature selection, which facilitates irrelevant feature exclusion and remarkable feature domination.

3. The Methodology: ELC

3.1. Model Description

In this paper, we use $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ to denote the data set, where $\mathbf{X} = [\mathbf{x}_1; ...; \mathbf{x}_n] \in \mathbb{R}^{n\times d}$ represents the instance matrix and instances are characterized by d features in the feature set $\mathbf{F} = \{\mathbf{f}_1, ..., \mathbf{f}_d\}$. $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_l] \in \{0, 1\}^{n\times l}$ denotes the target label matrix, where $y_{ij} = 1$ represents a positive label and $y_{ij} = 0$ corresponds to a negative one.

Then, we formulate the multi-label feature selection by embedding label correlation (ELC) as follows:

$$\min_{\mathbf{W}} \frac{1}{2} \left\| \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2, \; s.t. \; \hat{\mathbf{Y}} = \frac{1}{n} (\mathbf{XW})^T \mathbf{Y}, \mathbf{W} \in \{0, 1\}^{d\times l}, \|\mathbf{W}\|_{2,0} = k, \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{l\times l}$ represents the label correlation matrix calculated over the initial label matrix, and k is the number of selected features. $\mathbf{W} \in \mathbb{R}^{d\times l}$ is the feature selection matrix, where w_{ij} indicates the importance (also known as weight) of the i-th feature to the j-th label.

Equation (1) is actually the feature evaluation function of ELC, which is essentially a Frobenius-norm quadratic model. The matrix $\mathbf{S}$ represents the label correlations extracted from the label space, and its each element describes a relation between two target labels. These correlations can be easily obtained by some quantitative measurements, including RBF kernel function, Pearson correlation coefficient, etc. $\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ represents the label correlations extracted from the reduced feature space. $\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ is differentiated from $\mathbf{S}$ on account of the disturbance of noisy features. As described in Section 1, noisy features may distort the structure of the feature space and provide negative learning information. Considering this, ELC evaluates features based on their abilities of preserving label correlations in the feature space, that is, keeping feature-label space consistency. The features that can minimize the discrepancy between $\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$ and $\mathbf{S}$ will be highly scored by ELC. In this way, ELC can be expected to construct an optimal feature subspace with eliminating different kinds of noisy features.

Under the constraint of the $\ell_{2,0}$-norm in Equation (1), only k row in $\mathbf{W}$ is nonzero. This corresponds to the k selected features for l target labels, where 1 represents selected and 0 represents none. Note that k is most likely to be unequal to l. That is, more than one feature may be selected responsible for discriminating the same label, or only one feature is discriminative for more than one label. In the former case, multiple features are unified to recognize one target, while one feature deals with multiple recognition sub-tasks in the latter case.

3.2. Property Analysis

The feature subset $\hat{\mathbf{F}} = \left\{\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \ldots, \hat{\mathbf{f}}_k\right\}$ that is selected by ELC can be considered as maximally maintaining feature-label space consistency. $\hat{\mathbf{F}}$ is expected to be constituted by the remarkable features and exclude the noisy ones. In this subsection, we will further analyze the properties of ELC and reveal its underlying characteristics.

Suppose that each feature in $\mathbf{F}$ has been standardized to have mean zero and unit length. Then, the following things hold for Equation (1):

$$\left\| \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2 = \left\| \frac{1}{n^2} \left(\mathbf{Y}^T(\mathbf{XW})(\mathbf{XW})^T\mathbf{Y} \right) - \mathbf{S} \right\|_F^2.$$

This is the objective of ELC. For more clearly illustrating its properties, let $\hat{\mathcal{S}} = n^2\mathbf{S}$ and $\mathcal{H} = \mathbf{Y}^T(\mathbf{XW})(\mathbf{XW})^T\mathbf{Y}$. Then,

$$\left\| \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - \mathbf{S} \right\|_F^2 = \frac{1}{n^2} \left(tr(\mathcal{H}^T\mathcal{H}) + tr(\hat{\mathcal{S}}^T\hat{\mathcal{S}}) - 2tr(\hat{\mathcal{S}}^T\mathcal{H}) \right).$$

Three terms are involved in this equation. Clearly, $tr(\hat{\mathcal{S}}^T\hat{\mathcal{S}})$ represents the label correlation information extracted from the label space and is constant in the selection process. Thus, it is easy to

conclude that $\min_{\mathbf{W}} \left\|\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - \mathbf{S}\right\|_F^2$ is equivalent to $\min_{\mathbf{W}} tr(\mathcal{H}^T\mathcal{H})$ and $\max_{\mathbf{W}} tr(\hat{\mathcal{S}}^T\mathcal{H})$. Then, two properties of ELC are given as follows:

Property 1. *Label correlation information can be maximally embedded in feature selection by ELC.*

Proof. $tr(\hat{\mathcal{S}}^T\mathcal{H}) = tr\left((\mathbf{XW})^T\mathbf{Y}\hat{\mathcal{S}}\mathbf{Y}^T(\mathbf{XW})\right) = \sum_{i=1}^{k}\hat{\mathbf{f}}_i^T(\mathbf{Y}\hat{\mathcal{S}}\mathbf{Y}^T)\hat{\mathbf{f}}_i = \sum_{i=1}^{k}\hat{\mathbf{f}}_i^T\left(\sum_{c1=1}^{l}\sum_{c2=1}^{l}\mathbf{y}_{c1}s_{c1,c2}\mathbf{y}_{c2}^T\right)\hat{\mathbf{f}}_i$, where $s_{c1,c2}$ is the correlation degree of the labels $\mathbf{y}_{c1}$ and $\mathbf{y}_{c2}$, and $\mathbf{XW}$ indicates the selected features. Then, the following things holds: $\min_{\mathbf{W}} \left\|\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - \mathbf{S}\right\|_F^2 \propto \max_{\mathbf{W}} \sum_{i=1}^{k}\hat{\mathbf{f}}_i^T\left(\sum_{c1=1}^{l}\sum_{c2=1}^{l}\mathbf{y}_{c1}s_{c1,c2}\mathbf{y}_{c2}^T\right)\hat{\mathbf{f}}_i$. $\sum_{c1=1}^{l}\sum_{c2=1}^{l}\mathbf{y}_{c1}s_{c1,c2}\mathbf{y}_{c2}^T$ can be regarded as the correlation information of pairwise labels. Therefore, ELC can maximally embed label correlations in its feature selection process. □

Label correlation information is important for multi-label learning. For example, the images about seas may share some common labels for recognition, such as ship, fish, and seagull, and their close correlations may help us distinguish the image category and find their shared features. The existing multi-label learning methods are categorized on the basis of the label correlation orders they consider [39]. Their correlation modeling capabilities directly affect their discriminative performance. As demonstrated in Property 1, ELC can measure the pairwise label correlations. Furthermore, it can also preserve this correlation information in its constructed feature subspace, which is crucial for ELC to eliminate noisy features. In other words, the features that can maximally preserve label correlation information are preferred by ELC. This strategy facilitates ELC building a low-dimensional feature space that is consistent with the label space and also suitable for multi-label learning.

In addition to the above property with respect to maximally embedding label correlations, another important property of ELC is illustrated as follows:

Property 2. *Feature redundancy can be minimized by ELC.*

Proof. $tr(\mathcal{H}^T\mathcal{H}) = \sum_{i,j=1}^{k}\left((\hat{\mathbf{f}}_i^T\mathbf{Y})(\hat{\mathbf{f}}_j^T\mathbf{Y})^T\right)^2 = \sum_{i,j=1}^{k}\sum_{c=1}^{l}\left(\langle\hat{\mathbf{f}}_i,\mathbf{y}_c\rangle\langle\hat{\mathbf{f}}_j,\mathbf{y}_c\rangle\right)^2$
$= \sum_{i,j=1}^{k}\sum_{c=1}^{l} n^4\sigma_{\mathbf{y}_c}^4\rho_{\hat{\mathbf{f}}_i,\mathbf{y}_c}^2\rho_{\hat{\mathbf{f}}_j,\mathbf{y}_c}^2$,
where $\sigma_{\mathbf{y}_c}$ is the standard deviation of the label $\mathbf{y}_c$, and $\rho_{\hat{\mathbf{f}}_i,\mathbf{y}_c}$ and $\rho_{\hat{\mathbf{f}}_j,\mathbf{y}_c}$ are the Pearson correlation coefficients of $\mathbf{y}_c$ with the features $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{f}}_j$, respectively. Then, we have $\min_{\mathbf{W}} \left\|\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} - \mathbf{S}\right\|_F^2 \propto \min_{\mathbf{W}} \sum_{i,j=1}^{k}\sum_{c=1}^{l} n^4\sigma_{\mathbf{y}_c}^4\rho_{\hat{\mathbf{f}}_i,\mathbf{y}_c}^2\rho_{\hat{\mathbf{f}}_j,\mathbf{y}_c}^2$.

Clearly, n and $\sigma_{\mathbf{y}_c}$ are constant in the feature selection process. $\sum_{c=1}^{l}\rho_{\hat{\mathbf{f}}_i,\mathbf{y}_c}\rho_{\hat{\mathbf{f}}_j,\mathbf{y}_c}$ can be regarded as the shared label dependency of the features $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{f}}_j$, that is, the feature redundancy for recognizing the target $\mathbf{y}_c$. Therefore, ELC can minimize feature redundancy in its feature selection process. □

Note that the term $\sum_{c=1}^{l}\rho_{\hat{\mathbf{f}}_i,\mathbf{y}_c}\rho_{\hat{\mathbf{f}}_j,\mathbf{y}_c}$ in Property 2 is obtained by introducing the label correlation information. This is a completely novel estimation for the label-specific feature redundancy. The most majority of existing feature selection approaches (including the single-label and multi-label ones) adopt a univariate measurement criterion and merely the top-k features have opportunities to prevail. This strategy largely increases the redundant recognition information shared between features. For example, if we select the genes that are all discriminative for the diabetes type 1, we probably cannot give an accurate diagnosis since these features may be less aware of other types of diabetes. This is why we have to reduce recognition redundancy and enrich recognition information. Some approaches are able to reduce feature redundancy, while their focus is not the label-specific redundancy. For example, $\sum_{i,j=1}^{k}\rho_{\hat{\mathbf{f}}_i,\hat{\mathbf{f}}_j}$ is actually reduced in SPFS [20]. This term includes an additional information irrelevant to recognition, and correspondingly, it is inappropriate. In contrast, ELC removes label-specific feature redundancy and is more suitable for multi-label learning with eliminating noisy features.

As discussed above, ELC processes two properties, i.e., maximally preserving label correlation information and minimizing label-specific feature redundancy. These characteristics account for the superior ability of ELC in eliminating noisy features and picking out remarkable ones.

4. Multi-Task Optimization for ELC

Equation (1) describes an integer programming problem, which is NP-hard and complicated to solve. Moreover, the $\ell_{2,0}$-norm constraint in Equation (1) is non-smooth, which leads to a slow convergence rate. In this section, we devise an efficient framework to address this problem by using the sparse multi-task learning technology in the proximal alternating direction method (PADM) framework [40].

Suppose the spectral decomposition of the correlation matrix $\mathbf{S}$ can be denoted as

$$\mathbf{S} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^T = \mathbf{\Phi} diag(\sigma_1, \ldots, \sigma_l)\mathbf{\Phi}^T, \sigma_1 \geq \ldots \geq \sigma_l,$$

where $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ are respectively the eigenvector and eigenvalue matrices of $\mathbf{S}$. Then, Equation (1) can be reformulated as

$$\min_{\mathbf{W},\mathbf{p}} \frac{1}{2} \left\| \mathbf{Y}^T\mathbf{X} diag(\mathbf{p})\mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2, \; s.t. \; \mathbf{W} \in \mathbb{R}^{d\times l}, \|\mathbf{W}\|_{2,1} \leq t, \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = k, \tag{2}$$

where $\mathbf{\Gamma}^* = n\mathbf{\Phi}\mathbf{\Sigma}^{1/2}$, t is a hyperparameter to constrain $\|\mathbf{W}\|_{2,1}$ to a convex solution, $\mathbf{p}$ is a feature indicator vector that reflects whether the corresponding features are selected or not (1 for selected and 0 for otherwise), and $\mathbf{1}$ is the vector with all ones.

On the basis of Equation (2), ELC is actually reformulated as a multivariate regression problem, which enables the multi-task learning technology [41]. This technology aims to learn a common set of features to tackle multiple relevant tasks and excels at various sparse learning formulations, including the optimization problem in Equation (1). Based on the multi-task learning technology, we then obtain the equivalent form of ELC as follows:

$$\min_{\mathbf{W},\mathbf{p}} \frac{1}{2} \left\| \hat{\mathbf{A}} diag(\mathbf{p})\mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2 + \lambda \|\mathbf{W}\|_{2,1}, \; s.t. \; \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = k, \tag{3}$$

where $\hat{\mathbf{A}} = \mathbf{Y}^T\mathbf{X}$, and $\lambda > 0$ is the regularization parameter. Clearly, we can apply the augmented Lagrangian method to solve this problem. Then, Equation (3) is further reformulated as

$$\min_{\mathbf{U},\mathbf{W},\mathbf{p}} \frac{1}{2} \left\| \hat{\mathbf{A}} diag(\mathbf{p})\mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2 + \lambda \|\mathbf{U}\|_{2,1}, \; s.t. \; \mathbf{U} = \mathbf{W}, \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = k. \tag{4}$$

The Lagrangian function can be defined as

$$\mathcal{L}(\mathbf{U},\mathbf{W},\mathbf{p},\mathbf{V}) = \frac{1}{2} \left\| \hat{\mathbf{A}} diag(\mathbf{p})\mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2 + \frac{\beta}{2} \|\mathbf{W} - \mathbf{U}\|^2 + \lambda \|\mathbf{U}\|_{2,1} - tr\left(\mathbf{V}^T(\mathbf{W} - \mathbf{U})\right), \tag{5}$$

where $\mathbf{V} = \left(\mathbf{v}_1^T, \ldots, \mathbf{v}_d^T\right)^T \in \mathbb{R}^{d\times l}$ is the Lagrangian multiplier, and $\beta > 0$ is the penalty parameter.

Equation (5) involves four variables, that is, the auxiliary variable $\mathbf{U}$, the feature weight matrix $\mathbf{W}$, the feature indicator vector $\mathbf{p}$, and the Lagrangian multiplier $\mathbf{V}$. Clearly, simultaneously optimizing four variables is impractical. Accordingly, $\mathbf{V}$ is temporarily fixed for simplification in the following analysis. Then, minimizing $\mathcal{L}(\mathbf{U},\mathbf{W},\mathbf{p},\mathbf{V})$ is equivalent to the following two subproblems; i.e.,

- $\min_{\mathbf{U}} \mathcal{L}_1(\mathbf{U}) = \min_{\mathbf{U}} \frac{\beta}{2} \|\mathbf{W} - \mathbf{U}\|^2 + \lambda \|\mathbf{U}\|_{2,1} + tr(\mathbf{V}^T\mathbf{U})$;
- $\min_{\mathbf{W},\mathbf{p}} \mathcal{L}_2(\mathbf{W},\mathbf{p}) = \min_{\mathbf{W},\mathbf{p}} \frac{1}{\beta} \left\| \hat{\mathbf{A}} diag(\mathbf{p})\mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2 + \|\mathbf{W} - \mathbf{U}\|^2 - \frac{2}{\beta} tr(\mathbf{V}^T\mathbf{W})$.

As to $\mathcal{L}_1(\mathbf{U})$, the following holds:

$$\mathcal{L}_1(\mathbf{U}) = \sum_{i=1}^{d} \left(\frac{\beta}{2} \left\| \mathbf{w}^i - \mathbf{u}^i \right\|^2 + \lambda \left\| \mathbf{u}^i \right\| + tr(\mathbf{v}_i^T \mathbf{u}^i) \right), \tag{6}$$

where $\mathbf{w}^i$ and $\mathbf{u}^i$ are the i-th row vectors of $\mathbf{W}$ and $\mathbf{U}$, respectively. Then, we reformulate $\min_{\mathbf{U}} \mathcal{L}_1(\mathbf{U})$ to its close form [41] as

$$\min_{\mathbf{u}^i} \sum_{i=1}^{d} \left(\frac{\beta}{2} \left\| \mathbf{w}^i - \mathbf{u}^i + \frac{1}{\beta} \mathbf{v}_i \right\|^2 + \lambda \left\| \mathbf{u}^i \right\| \right). \tag{7}$$

Conducting gradient descent on Equation (7) yields the following optimal solution as

$$\mathbf{u}^i = \max \left\{ \left\| \mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i \right\| - \frac{\lambda}{\beta}, 0 \right\} \frac{\mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i}{\left\| \mathbf{w}^i + \frac{1}{\beta} \mathbf{v}_i \right\|}. \tag{8}$$

Then, the optimal $\mathbf{U}$ in iteration $[t+1]$ can be denoted as

$$\mathbf{U}^{[t+1]} = \max \left\{ \left\| \mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]} \right\| - \frac{\lambda}{\beta}, 0 \right\} \frac{\mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]}}{\left\| \mathbf{W}^{[t]} + \frac{1}{\beta} \mathbf{V}^{[t]} \right\|}. \tag{9}$$

In terms of $\min_{\mathbf{W},\mathbf{p}} \mathcal{L}_2(\mathbf{W}, \mathbf{p})$, we let $\mathcal{P} = \{\mathbf{p} | \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T \mathbf{1} = k\}$. The dual problem of $\min_{\mathbf{W},\mathbf{p}} \mathcal{L}_2(\mathbf{W}, \mathbf{p})$ is

$$\min_{\mathbf{p} \in \mathcal{P}} \max_{\mathbf{W}} \mathcal{L}_2(\mathbf{W}, \mathbf{p}). \tag{10}$$

Since simultaneously solving the both variables $\mathbf{p}$ and $\mathbf{W}$ is still tough, we first fix $\mathbf{p}$ to optimize $\mathbf{W}$. Then, the solution of $\mathbf{W}$ can be obtained as

$$\left(diag(\mathbf{p}) \hat{\mathbf{A}}^T \hat{\mathbf{A}} diag(\mathbf{p}) - \beta \mathbf{I} \right) \mathbf{W} = diag(\mathbf{p}) \hat{\mathbf{A}}^T \mathbf{\Gamma}^* + \beta \mathbf{U} + \mathbf{V}, \tag{11}$$

where $\mathbf{I}$ is the identity matrix. The structure of $\hat{\mathbf{A}}^T \hat{\mathbf{A}}$ is commonly not circulant, and therefore the computation of Equation (11) is involved [42]. Considering this, an approximate term is added to $\mathcal{L}_2(\mathbf{W}, \mathbf{p})$ as follows:

$$\begin{aligned} \tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p}) &= \frac{1}{\beta\tau} \left\| \mathbf{W} - \mathbf{W}^{[t]} + \tau \mathbf{\Omega}^{[t]} \right\| - \frac{2}{\beta} tr(\mathbf{V}^T \mathbf{W}) + \| \mathbf{W} - \mathbf{U} \|^2, \\ \mathbf{\Omega}^{[t]} &= diag(\mathbf{p}^{[t]}) \hat{\mathbf{A}}^T \left(\hat{\mathbf{A}} diag(\mathbf{p}^{[t]}) \mathbf{W}^{[t]} - \mathbf{\Gamma}^* \right), \end{aligned} \tag{12}$$

where $\tau > 0$, and $\mathbf{W}^{[t]}$ is the optimal value of $\mathbf{W}$ in iteration $[t]$. Then, the solution of $\mathbf{W}^{[t+1]}$ is

$$\mathbf{W}^{[t+1]} = \left(\frac{\tau}{\beta\tau + 1} \right) \left(\beta \mathbf{U}^{[t+1]} + \mathbf{V}^{[t]} + \frac{1}{\tau} (\mathbf{W}^{[t]} - \tau \mathbf{\Omega}^{[t]}) \right). \tag{13}$$

The detailed inference can be found in the Appendix A.

Similarly, we can easily obtain the optimal $\mathbf{p}$ by fixing $\mathbf{W}$. Equation (10) is then equivalent to the following minimization problem in this case as follows:

$$\min_{\mathbf{p} \in \mathcal{P}} \left\| \hat{\mathbf{A}} diag(\mathbf{p}) \mathbf{W} - \mathbf{\Gamma}^* \right\|_F^2 = \min_{\mathbf{p} \in \mathcal{P}} \left\| \mathbf{Y}^T \sum_{i=1}^{d} p_i \mathbf{f}_i \mathbf{w}^i - \mathbf{\Gamma}^* \right\|_F^2. \tag{14}$$

Apparently, the top-k features that minimize $\left\| \mathbf{Y}^T \mathbf{f}_i \mathbf{w}^i - \mathbf{\Gamma}^* \right\|_F^2$ can be regarded as the remarkable ones. Their corresponding values in $\mathbf{p}$ are assigned as 1.

Note that the Lagrangian multiplier $\mathbf{V}$ is fixed through the above analysis, mainly for simplifying the solution process. We further tackle this problem in the popular PADM framework as illustrated in Algorithm 1. In this framework, $\mathbf{V}$ can be updated as

$$\mathbf{V}^{[t+1]} = \mathbf{V}^{[t]} - \beta\left(\mathbf{W}^{[t+1]} - \mathbf{U}^{[t+1]}\right). \tag{15}$$

Algorithm 1 ELC.

input: $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_d\}, \mathbf{Y}, \mathbf{S}, k, \beta, \tau, \lambda$

output: $\mathbf{p}^{[t]}$

1: **begin**

2: $t = 0, \mathbf{W}^{[0]} = \mathbf{0}_{d\times l}, \mathbf{U}^{[0]} = \mathbf{0}_{d\times l}, \mathbf{V}^{[0]} = \frac{1}{d}\mathbf{1}_{d\times l}$;

3: find top-k features $\hat{\mathbf{f}}_1^{[0]}, \ldots, \hat{\mathbf{f}}_k^{[0]}$ that minimize Equation (1), and set $p_i^{[0]} = \begin{cases} 1, \ \mathbf{f}_i \in \left\{\hat{\mathbf{f}}_1^{[0]}, \ldots, \hat{\mathbf{f}}_k^{[0]}\right\} \\ 0, \ otherwise \end{cases}$;

4: **while** "not converged" **do**

5: optimize $\mathbf{U}^{[t+1]}$ according to Equation (9);

6: optimize $\mathbf{W}^{[t+1]}$ according to Equation (13);

7: find top-k features $\hat{\mathbf{f}}_1^{[t+1]}, \ldots, \hat{\mathbf{f}}_k^{[t+1]}$ which minimize Equation (14), and set $p_i^{[t+1]} = \begin{cases} 1, \ \mathbf{f}_i \in \left\{\hat{\mathbf{f}}_1^{[t+1]}, \ldots, \hat{\mathbf{f}}_k^{[t+1]}\right\} \\ 0, \ otherwise \end{cases}$;

8: update $\mathbf{V}^{[t+1]}$ according to Equation (15);

9: $t = t + 1$;

10: **end while**;

11: **return** $\mathbf{p}^{[t]}$;

12: **end;**

ELC in Algorithm 1 is implemented in the regression framework PADM, which is a fast alternating approach for the well-known alternating direction method (ADM) framework. PADM is effective and efficient in solving the minimization problem of the augmented Lagrangian function, and is able to converge to a certain solution $\{\mathbf{W}^*, \mathbf{U}^*\}$ from any starting point $\left\{\mathbf{W}^{[0]}, \mathbf{U}^{[0]}\right\}$ for any $\beta > 0$ [40].

In terms of the complexity of ELC, it only takes $O(k \log d)$ time to find k remarkable features from the d candidates. Thus, the time consumption for line 3 is $O(ndl^2 + k \log d)$. The cost of the while loop in Algorithm 1 mainly lies in lines 6 and 7, which is $O(d^2l^2 + ndl^2 + k \log d)$. As this iteration process is repeated for t times, its total cost is $O(t(d^2l^2 + ndl^2 + k \log d))$. Suppose $t \gg 1$. Then, the total complexity of ELC is approximately equal to $O(t(d^2l^2 + ndl^2 + k \log d))$, where d, n, l, k, t are the numbers of features, instances, labels, selected features, and iterations for convergence, respectively.

5. Experimental Evaluation

Fourteen groups of multi-label data sets fetched from the Mulan library (http://mulan.sourceforge.net/datasets-mlc.html) are taken as the benchmarks in this section, which are shown in Table 1. We compare ELC (the source code is available at https://github.com/wangjuncs/ELC) with the following state-of-the-art multi-label feature selection methods:

- MIFS (multi-label informed feature selection) [33]: a label correlation-based multi-label feature selection approach, which maps label information into a low-dimensional subspace and captures the correlations among multiple labels;

- CMFS (correlated and multi-label feature selection) [35]: a feature selection approach based on non-negative matrix factorization, which exploits the label correlation information in features, labels, and instances to select the relevant features and remove the noisy ones;
- LLSF (learning label-specific features) [36]: a unified multi-label learning framework for both feature selection and classification, which models high-order label correlations to select label-specific features.

Table 1. Benchmarks for multi-label feature selection.

Data Set	♯Features	♯Instances	♯Labels	Domain
emotions	72	539	6	music
yeast	103	2417	14	biology
birds	260	645	19	audio
enron	1001	1702	53	text
genbase	1186	662	27	biology
business	21,924	11,214	30	text
arts	23146	7484	26	text
education	27,534	12,030	33	text
reaction	30,324	12,828	22	text
health	30,605	9205	32	text
computers	34,096	12,444	33	text
science	37,187	6428	40	text
reference	39,679	8027	33	text
society	49,060	14,512	22	text

More detailed experimental configurations can be found in the Appendix B.

5.1. Example 1: Classification Performance

The average classification performance of each feature selection approach is recorded in Table 2 and the pairwise *t*-tests at 5% significance level were conducted to validate the statistical significance. In addition to the traditional precision and AUC metrics, hamming loss penalizes incorrect the recognitions of instances to each target label, ranking loss penalizes the misordered labels in pairs, and one-error penalizes the instances whose top-ranked predicted labels are not in the ground-truth label set. Five metrics evaluated the multi-label classification performance from different aspects.

A single metric is insufficient to illustrate the general classification performance on a dataset. For example, the overall performance of ML-KNN classifier [43] on birds is worse than that on enron under the precision metric, while it shows a better performance on birds than on enron under the AUC metric. Therefore, we extensively used five metrics to compare the performances of the compared approaches. As shown in Table 2, ELC outperforms MIFS, CMFS, and LLSF under various metrics. This superiority is attributed to two reasons. That is, ELC can effectively eliminate noisy features from the candidate feature subsets and maximally embed label correlation information into its selection process. The first term rules out the selection disturbance in the feature space, and the second term promises the proper guiding information extracted from the label space. By seamlessly fusing these two terms, ELC is able to find discriminative features for the downstream learning tasks. This point will be further validated in Sections 5.2 and 5.3.

Table 2. Average multi-label classification performance (mean ± std.): the best results and those not significantly worse than it are highlighted in bold (pairwise *t*-test at 5% significance level).

(a) Precision (the higher the better).

Approaches	Data Sets							*AVG.*
	Emotions	**Yeast**	**Birds**	**Enron**	**Genbase**	**Business**	**Arts**	
MIFS	0.6667 ± 0.04	0.7520 ± 0.02	0.3938 ± 0.02	0.6139 ± 0.03	0.7361 ± 0.15	0.8812 ± 0.00	0.5108 ± 0.01	*0.6506*
CMFS	0.7221 ± 0.02	0.7464 ± 0.01	0.4116 ± 0.03	0.6206 ± 0.01	0.7342 ± 0.15	0.892 ± 0.00	0.5611 ± 0.00	*0.6697*
LLSF	0.7016 ± 0.02	0.7532 ± 0.01	0.4231 ± 0.07	0.6197 ± 0.04	0.7352 ± 0.15	0.8924 ± 0.00	0.5615 ± 0.01	*0.6695*
ELC	**0.7306 ± 0.02**	**0.7564 ± 0.01**	**0.4671 ± 0.08**	**0.6347 ± 0.01**	**0.9868 ± 0.00**	**0.8931 ± 0.00**	**0.5646 ± 0.00**	***0.7190***
	Education	**Reaction**	**Health**	**Computers**	**Science**	**Reference**	**Society**	
MIFS	0.5129 ± 0.01	0.5836 ± 0.01	0.7391 ± 0.02	0.6629 ± 0.01	0.4592 ± 0.02	0.6267 ± 0.01	0.5899 ± 0.01	*0.5963*
CMFS	0.6164 ± 0.01	0.5883 ± 0.01	0.7437 ± 0.01	0.6931 ± 0.00	0.5477 ± 0.01	0.6718 ± 0.01	0.6463 ± 0.01	*0.6439*
LLSF	0.6163 ± 0.01	0.5880 ± 0.01	0.7435 ± 0.01	0.6932 ± 0.00	0.5478 ± 0.01	0.6720 ± 0.00	0.6460 ± 0.01	*0.6438*
ELC	**0.6213 ± 0.00**	**0.5952 ± 0.00**	**0.7469 ± 0.01**	**0.6962 ± 0.00**	**0.5565 ± 0.00**	**0.6742 ± 0.00**	**0.6500 ± 0.00**	***0.6486***

(b) AUC (the higher the better).

Approaches	Data Sets							*AVG.*
	Emotions	**Yeast**	**Birds**	**Enron**	**Genbase**	**Business**	**Arts**	
MIFS	0.6601 ± 0.53	0.6554 ± 0.03	0.6497 ± 0.02	0.5968 ± 0.03	0.7886 ± 0.10	0.6371 ± 0.01	0.6100 ± 0.01	*0.6568*
CMFS	0.7307 ± 0.03	0.6473±0.03	0.6403 ± 0.04	0.6194 ± 0.01	0.7883 ± 0.10	0.6821 ± 0.01	0.6606 ± 0.01	*0.6812*
LLSF	0.7069 ± 0.02	0.6601 ± 0.02	0.6793 ± 0.05	0.6092 ± 0.03	0.7887 ± 0.10	0.6824 ± 0.01	0.6608 ± 0.01	*0.6839*
ELC	**0.7513 ± 0.02**	**0.6706 ± 0.02**	**0.7018 ± 0.06**	**0.6385 ± 0.01**	**0.9663 ± 0.00**	**0.6834 ± 0.00**	**0.6659 ± 0.00**	***0.7254***
	Education	**Reaction**	**Health**	**Computers**	**Science**	**Reference**	**Society**	
MIFS	0.5830 ± 0.02	0.7065 ± 0.01	0.6994 ± 0.01	0.6364 ± 0.01	0.6109 ± 0.02	0.6246 ± 0.01	0.5964 ± 0.01	*0.6423*
CMFS	0.6753 ± 0.00	0.7111 ± 0.01	0.7014 ± 0.01	0.6864 ± 0.01	0.6732 ± 0.01	0.6674 ± 0.01	0.6430 ± 0.01	*0.6867*
LLSF	0.6761 ± 0.01	0.7114 ± 0.01	0.7032 ± 0.01	0.6859 ± 0.01	0.6723 ± 0.01	0.6672 ± 0.01	0.6427 ± 0.01	*0.6867*
ELC	**0.6779 ± 0.00**	**0.7188 ± 0.01**	**0.7053 ± 0.01**	**0.6905 ± 0.00**	**0.6789 ± 0.00**	**0.6681 ± 0.01**	**0.6465 ± 0.00**	***0.6911***

Table 2. *Cont.*

(c) Hamming loss (the lower the better).

Approaches	Data Sets							*AVG.*
	Emotions	**Yeast**	**Birds**	**Enron**	**Genbase**	**Business**	**Arts**	
MIFS	0.2865 ± 0.02	0.2006 ± 0.01	0.0538 ± 0.00	0.0544 ± 0.00	0.0303 ± 0.02	0.0270 ± 0.00	0.0610 ± 0.00	*0.1019*
CMFS	0.2600 ± 0.01	0.2031 ± 0.01	0.0535 ± 0.00	0.0527 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00	*0.0974*
LLSF	0.2697 ± 0.01	0.1999 ± 0.01	0.0532 ± 0.00	0.0539 ± 0.00	0.0303 ± 0.02	0.0253 ± 0.00	0.0568 ± 0.00	*0.0984*
ELC	**0.2517 ± 0.02**	**0.1982 ± 0.01**	**0.0518 ± 0.00**	**0.0523 ± 0.00**	**0.0049 ± 0.00**	**0.0251 ± 0.00**	**0.0567 ± 0.00**	***0.0915***
	Education	**Reaction**	**Health**	**Computers**	**Science**	**Reference**	**Society**	
MIFS	0.0430 ± 0.00	0.0539 ± 0.00	0.0373 ± 0.00	0.0379 ± 0.00	0.0353 ± 0.00	0.0317 ± 0.00	0.0557 ± 0.00	*0.0399*
CMFS	0.0371 ± 0.00	0.0536 ± 0.00	0.0368 ± 0.00	0.0350 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0511 ± 0.00	*0.0394*
LLSF	0.0371 ± 0.00	0.0536 ± 0.00	0.0369 ± 0.00	0.0349 ± 0.00	0.0322 ± 0.00	0.0280 ± 0.00	0.0512 ± 0.00	*0.0394*
ELC	**0.0368 ± 0.00**	**0.0531 ± 0.00**	**0.0366 ± 0.00**	**0.0348 ± 0.00**	**0.0319 ± 0.00**	**0.0278 ± 0.00**	**0.0508 ± 0.00**	***0.0391***

(d) Ranking loss (the lower the better).

Approaches	Data Sets							*AVG.*
	Emotions	**Yeast**	**Birds**	**Enron**	**Genbase**	**Business**	**Arts**	
MIFS	0.3154 ± 0.05	0.1767 ± 0.01	0.2988 ± 0.01	0.0986 ± 0.01	0.0586 ± 0.03	0.0377 ± 0.00	0.1510 ± 0.05	*0.1624*
CMFS	0.2404 ± 0.02	0.1810 ± 0.01	0.2887 ± 0.02	0.0959 ± 0.00	0.0594 ± 0.03	0.0336 ± 0.00	0.1341 ± 0.00	*0.1476*
LLSF	0.2711 ± 0.02	0.1756 ± 0.01	0.2777 ± 0.05	0.0962 ± 0.01	0.0591 ± 0.03	0.0335 ± 0.00	0.1341 ± 0.00	*0.1496*
ELC	**0.2379 ± 0.02**	**0.1733 ± 0.01**	**0.2568 ± 0.05**	**0.0924 ± 0.00**	**0.0065 ± 0.00**	**0.0334 ± 0.00**	**0.1332 ± 0.00**	***0.1334***
	Education	**Reaction**	**Health**	**Computers**	**Science**	**Reference**	**Society**	
MIFS	0.0988 ± 0.00	0.1459 ± 0.01	0.0498 ± 0.00	0.0820 ± 0.00	0.1351 ± 0.01	0.0849 ± 0.00	0.1372 ± 0.00	*0.0994*
CMFS	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0735 ± 0.00	0.1112 ± 0.00	0.0734 ± 0.00	0.1175 ± 0.00	*0.0897*
LLSF	0.0768 ± 0.00	0.1438 ± 0.01	0.0492 ± 0.00	0.0736 ± 0.00	0.1111 ± 0.00	0.0736 ± 0.00	0.1176 ± 0.00	*0.0898*
ELC	**0.0759 ± 0.00**	**0.1405 ± 0.00**	**0.0486 ± 0.00**	**0.0728 ± 0.00**	**0.1085 ± 0.00**	**0.0731 ± 0.00**	**0.1161 ± 0.00**	***0.0883***

Table 2. *Cont.*

(e) One error (the lower the better).

Approaches	Data Sets							*AVG.*
	Emotions	**Yeast**	**Birds**	**Enron**	**Genbase**	**Business**	**Arts**	
MIFS	0.4451 ± 0.05	0.2403 ± 0.01	0.7226 ± 0.03	0.3230 ± 0.03	0.3698 ± 0.21	0.1187 ± 0.00	0.6215 ± 0.01	*0.4059*
CMFS	0.3871 ± 0.02	0.2445 ± 0.01	0.6968 ± 0.04	0.3093 ± 0.02	0.3719 ± 0.21	0.1061 ± 0.00	0.5518 ± 0.01	*0.3811*
LLSF	0.3986 ± 0.03	0.2387 ± 0.01	0.6879 ± 0.08	0.3158 ± 0.05	0.3707 ± 0.21	0.1058 ± 0.00	0.5509 ± 0.01	*0.3812*
ELC	**0.3664 ± 0.03**	**0.2361 ± 0.01**	**0.6192 ± 0.11**	**0.2988 ± 0.01**	**0.0123 ± 0.00**	**0.1050 ± 0.00**	**0.5464 ± 0.00**	***0.3120***
	Education	**Reaction**	**Health**	**Computers**	**Science**	**Reference**	**Society**	
MIFS	0.6452 ± 0.02	0.5292 ± 0.02	0.3335 ± 0.03	0.4041 ± 0.01	0.6761 ± 0.02	0.4743 ± 0.01	0.4684 ± 0.01	*0.5104*
CMFS	0.4989 ± 0.01	0.5234 ± 0.02	0.3266 ± 0.02	0.3738 ± 0.01	0.5605 ± 0.02	0.4208 ± 0.01	0.3933 ± 0.01	*0.4537*
LLSF	0.4993 ± 0.01	0.5237 ± 0.02	0.3267 ± 0.02	0.3735 ± 0.00	0.5605 ± 0.02	0.4202 ± 0.01	0.3937 ± 0.01	*0.4536*
ELC	**0.4923 ± 0.00**	**0.5147 ± 0.01**	**0.3223 ± 0.02**	**0.3699 ± 0.00**	**0.5490 ± 0.01**	**0.4172 ± 0.00**	**0.3885 ± 0.00**	***0.4771***

5.2. Example 2: Eliminating Noisy Features

In this section, we evaluate the performances of the compared approaches in eliminating noisy features. We take emotions, birds, and enron as the benchmarks, and measure the residual feature redundancy in the selected feature subset $\hat{\mathbf{F}}$ as follows:

$$R(\hat{\mathbf{F}}) = \frac{1}{k'(k'-1)l} \sum_{\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j \in \hat{\mathbf{F}}} \sum_{c=1}^{l} \rho^2_{\hat{\mathbf{f}}_i, \mathbf{y}_c} \rho^2_{\hat{\mathbf{f}}_j, \mathbf{y}_c}, \tag{16}$$

where $\rho_{\hat{\mathbf{f}}_i, \mathbf{y}_c}$ and $\rho_{\hat{\mathbf{f}}_j, \mathbf{y}_c}$ are the Pearson correlation coefficients of the features $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{f}}_j$ with the target label $\mathbf{y}_l$, and k' and l are the numbers of the selected features and labels, respectively. When $R(\hat{\mathbf{F}})$ reaches its maximum value, the maximal redundant information exists in $\hat{\mathbf{F}}$, which interprets as the inferior ability of the selection approach in removing noisy features.

The feature redundancy of k' selected features for each approach is demonstrated in Figure 1, where $k' \in \{d/10, 2d/10, \ldots, 9d/10\}$ and d is the total number of original features. It illustrates that ELC is superior in reducing feature redundancy. In other words, ELC can effectively remove redundant features in its multi-label feature selection process. This is one of the crucial factors leading to the excellent discriminative ability of ELC. It should be pointed out that in contrast to the case of single-label feature selection, eliminating noisy features has not received sufficient attention from existing multi-label feature selection approaches. While the issue of noisy features is an obstacle of yielding high selection performance not only for the single-label learning but also for the multi-label cases, we devised ELC to comprehensively tackle this problem. Moreover, the reduced feature redundancy in the majority of redundancy elimination-based approaches is not directly relevant to the target labels. In contrast, ELC quantitatively reduces target-relevant redundancy without any prior probability knowledge, which is conducive to its superiority in multi-label feature selection.

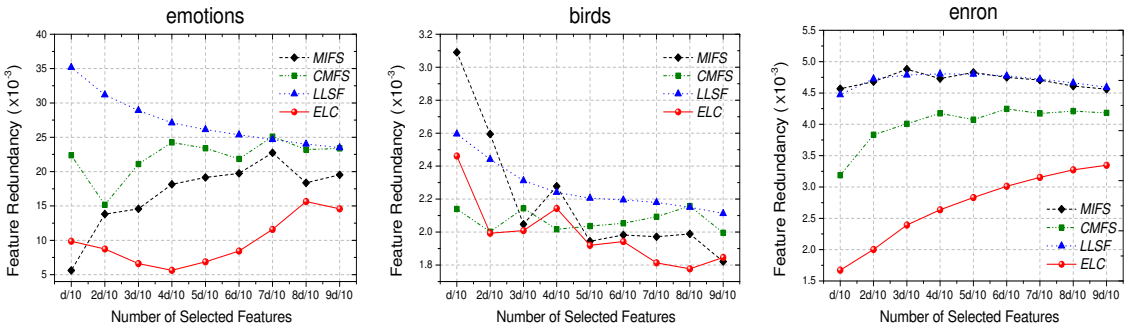

Figure 1. Classification redundancy: (**a**–**c**) are the classification redundancies produced by the feature selection approaches on the emotions, birds, and enron datasets, and the lower of the redundancy is the better.

5.3. Example 3: Embedding Label Correlations

Label correlation information is important for multi-label learning. In the following experiments, we estimate the preserved label correlation information of the selected feature subset $\hat{\mathbf{F}}$ as follows:

$$C(\hat{\mathbf{F}}) = \frac{1}{k'(k'-1)} \left\| \frac{1}{n^2} \mathbf{Y}^T \mathbf{X}_{\hat{\mathbf{F}}} \mathbf{X}^T_{\hat{\mathbf{F}}} \mathbf{Y} - \mathbf{S} \right\|^2_F, \tag{17}$$

where $\mathbf{X}_{\hat{\mathbf{F}}}$ denotes the instances characterized by $\hat{\mathbf{F}}$ and $\mathbf{S}$ is the label correlation matrix of the original data. Intuitively, Equation (17) measures the residue scale of label correlation information in the original

and reduced feature spaces. A lower value indicates more information preserved. In other words, more label correlation information can be embedded in the feature selection process in this situation.

Similarly to the configuration in Section 5.2, we take emotions, birds, and enron as the benchmarks and record $C(\hat{\mathbf{F}})$ of the k' features selected by each approach, where $k' \in \{d/10, 2d/10, \ldots, 9d/10\}$. As shown in Figure 2, ELC is better at preserving the class correlation information than the other multi-label feature selection approaches. Actually, the majority of the existing multi-label feature selection approaches take the label correlation information into consideration to some extent. In contrast to these approaches, ELC quantitatively measures this correlation information and maximally embeds it into the feature selection process. This characteristic, which has already been proved in Property 2, can be further revealed by the experimental results in this section.

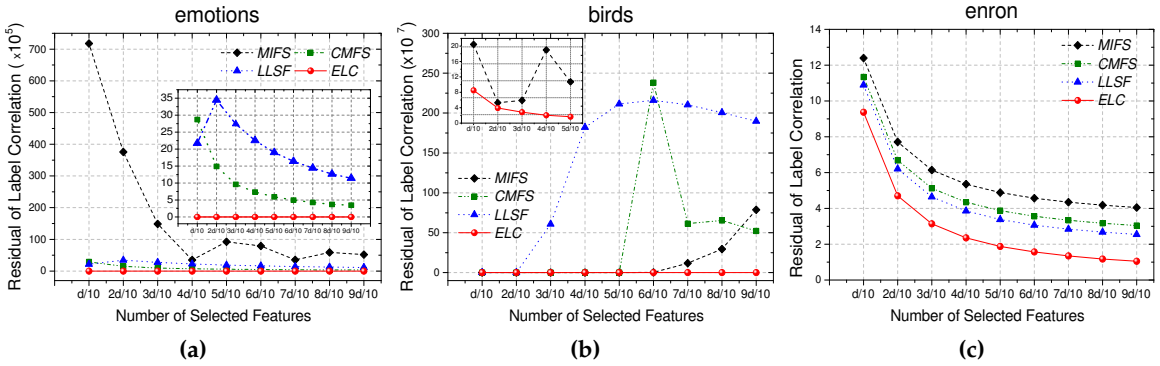

Figure 2. Residual label correlation information: (**a**–**c**) are the residual scales of the label correlation information that are not embedded by the feature selection approaches on the emotions, birds, and enron datasets, and the lower of the residual scale is the better.

5.4. Example 4: Time Consumption

In this section, we compare the approaches in terms of their feature selection efficiency. The time consumption here merely records the feature selection time, excluding the classification cost. All of the tests were implemented in Matlab on an Intel Core i7-4790 CPU (@3.6GHz) with 32GB memory (Intel Corp., Santa Clara, CA, USA). We respectively selected k' ($k' \in \{100, 300, 500, 700, 900\}$) features on the enron dataset and recorded the time consumption of each compared approach. As illustrated in Figure 3, ELC and CMFS are comparably efficient to converge, while MIFS is most time-consuming, which may be mainly attributed to its involved label clustering process.

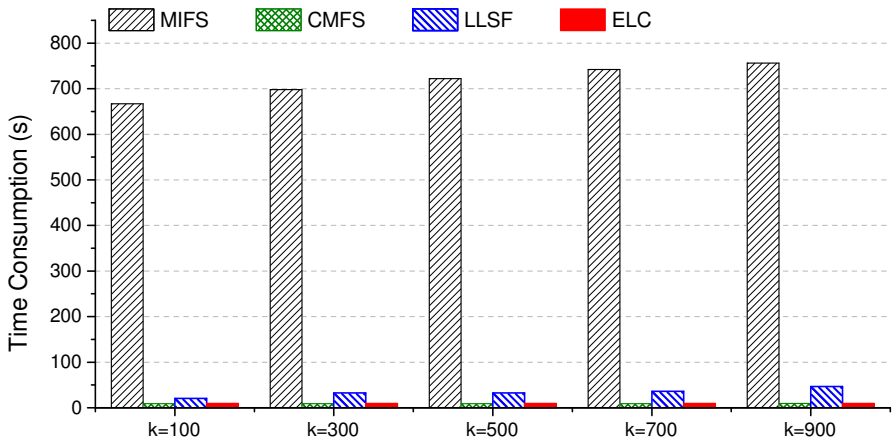

Figure 3. Time consumption of each multi-label feature selection approach on the enron dataset.

6. Conclusions

A novel multi-label feature selection method called ELC is proposed in this paper. ELC embeds label correlation information in reduced feature subspace to eliminate noisy features. In this way, irrelevant and redundant features can be expected to be removed and a discriminative feature subset is constructed for the downstream learning tasks. These advantages help ELC yield good feature selection performance on a wide broad of multi-label data sets under various evaluation metrics.

In terms of optimizing ELC, we can feed it to some gradient descent frameworks to efficiently yield its optimal values, such as Adam with a self-adaptive learning rate [44]. Another interesting and possible exploration would be the consideration of noisy labels, which would induce negative effects on estimating label correlations. According to our pilot study, noisy labels may distort the label space and provide inaccurate guide information for feature selection. How to eliminate noisy labels may inspire our future work.

Author Contributions: Each author greatly contributed to the preparation of this manuscript. J.W. (Jun Wang) and J.W. (Jinmao Wei) wrote the paper; Y.X. and H.X. designed and performed the experiments; Z.S. and Z.Y. devised the optimization algorithms. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (number 61772288), the Natural Science Foundation of Tianjin City (number 18JCZDJC30900), the Ministry of Education of Humanities and Social Science Project (number 16YJC790123), the National Natural Science Foundation of Shandong Province (number ZR2019MA049), and the Cooperative Education Project of the Ministry of Education of China (number 201902199006).

Acknowledgments: The authors are very grateful to the anonymous reviewers and editor for their helpful and constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

After adding an approximate term to $\mathcal{L}_2(\mathbf{W}, \mathbf{p})$ and reformulating it to $\tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p})$, we take the derivative of $\tilde{\mathcal{L}}_2(\mathbf{W}, \mathbf{p})$ with respect to $\mathbf{W}$ as follows:

$$\frac{\partial \tilde{\mathcal{L}}_2}{\partial \mathbf{W}} = \beta(\mathbf{W} - \mathbf{U}) - \mathbf{V} + \frac{1}{\tau}(\mathbf{W} - \mathbf{W}^{[t]} + \tau\mathbf{\Omega}^{[t]}), \mathbf{\Omega}^{[t]} = diag(\mathbf{p}^{[t]})\hat{\mathbf{A}}^T \left(\hat{\mathbf{A}} diag(\mathbf{p}^{[t]})\mathbf{W}^{[t]} - \mathbf{\Gamma}^*\right).$$

To induce the optimal solution of $\mathbf{W}$, we make $\frac{\partial \tilde{\mathcal{L}}_2}{\partial \mathbf{W}}$ equal to 0 and obtain:

$$(\beta + \frac{1}{\tau})\mathbf{W} = \beta\mathbf{U} + \mathbf{V} + \frac{1}{\tau}(\mathbf{W}^{[t]} - \tau\mathbf{\Omega}^{[t]}).$$

Then, the optimal solution of $\mathbf{W}$ in the iteration $[t+1]$ can be represented as

$$\mathbf{W}^{[t+1]} = \left(\frac{\tau}{\beta\tau + 1}\right)\left(\beta\mathbf{U}^{[t+1]} + \mathbf{V}^{[t]} + \frac{1}{\tau}(\mathbf{W}^{[t]} - \tau\mathbf{\Omega}^{[t]})\right).$$

Appendix B. Experimental Configuration

The correlation (or similarity) matrices involved in experiments are all calculated based on the RBF kernel function. Specifically, the label correlation matrix $\mathbf{S}$ in ELP is defined as $\mathbf{S}_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\delta'^2}\right), \langle \mathbf{y}_i, \mathbf{y}_j \rangle \neq 0 \\ 0, \ otherwise \end{cases}$, where $\delta'^2 = mean(\|\mathbf{y}_i - \mathbf{y}_j\|^2), i, j = 1, \ldots, l$. The instance similarity matrix in SPFS and CMFS is calculated as $\mathbf{K}_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\delta^2}\right), \mathbf{y}_i = \mathbf{y}_j \\ 0, \ otherwise \end{cases}$, where $\delta^2 = mean(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$. The affinity graph in MIFS is constructed as $\mathbf{K}_{ij} =$

$\begin{cases} \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\delta^2}\right), \ \mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}_j) \ or \ \mathbf{x}_j \in \mathcal{N}_p(\mathbf{x}_i) \\ 0; \ otherwise \end{cases}$, where $\mathcal{N}_p(\mathbf{x}_i)$ is the p-nearest neighbor of instance $\mathbf{x}_i$.

SPFS is implemented via the sequential forward selection (SFS) strategy. For a fair comparison, we tune the regularization parameter for all approaches via a grid search from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$. For ELC, the parameter β is fixed to $\beta = 10^8$, and τ is set to the spectral radius of $\hat{\mathbf{A}}^T\hat{\mathbf{A}}$ in the initial state and updated as $\tau^{[t]} = \frac{1}{\max(\|\psi^i\|)}$ in the t-th iteration, where ψ^i is the i-th row vector of $\mathbf{\Psi}$ and $\mathbf{\Psi} = \hat{\mathbf{A}}^T\hat{\mathbf{A}}\mathbf{V}^{[t]}$. The convergence state is reached when any of the following two conditions is satisfied: (1) $t_{max} = 10^3$; and (2) $\left\|\mathbf{W}^{[t+1]} - \mathbf{W}^{[t]}\right\| \leq 10^{-4}$.

Multi-label k-nearest neighbor (ML-kNN) classifier [43] is built on the k' features selected by each compared approach, when $k' \in \{d/10, 2d/10, \ldots, 9d/10\}$ and d is the total number of features. All of the numerical features are normalized to zero mean and unit variance, and we employ the excellent features selected by the compared approaches to construct the ML-kNN classifiers and compare their classification performances. The 5-fold cross-validation is conducted, and we report the average performance of the ML-kNN classification under five metrics, i.e., precision, AUC, Hamming loss, ranking loss, and one error [39].

References

1. Tang, J.; Alelyani, S.; Liu, H. Feature felection for classification: A review. In *Data Classification: Algorithms and Applications*; CRC Press: Chapman, CA, USA, 2014.
2. Wang, J.; Wei, J.; Yang, Z. Supervised feature selection by preserving class correlation. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 1613–1622.
3. Cai, D.; Zhang, C.; He, X. Efficient and robust feature selection via joint l2,1-norms minimization. In Proceedings of the KDD '10: The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 333–342.
4. Xu, Y.; Wang, J.; An, S.; Wei, J.; Ruan, J. Semi-supervised multi-label feature selection by preserving feature-label space consistency. In Proceedings of the CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Orino, Italy, 22–26 October 2018; pp. 783–792.
5. Brown, G.; Pocock, A.; Zhao, M.; Luján, M. Conditional Likelihood Maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.* **2012**, *12*, 27–66.
6. Gu, Q.; Li, Z.; Han, J. Generalized fisher score for feature selection. In Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 14–17 July 2011; pp. 266–273.
7. He, X.; Cai, D.; Niyogi, P. Laplacian score for feature selection. In Proceedings of the 18th International Conference on Neural Information Processing Systems, Shanghai, China, 13–17 November 2011; pp. 507–514.
8. Lin, D.; Tang, X. Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion. In Proceedings of the Computer Vision—ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 68–82.
9. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
10. Robnik-Šikonja, M.; Kononenko, I. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* **2003**, *53*, 23–69.
11. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
12. Bermejo, P.; Gámez, J.A.; Puerta, J.M. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. *Knowl.-Based Syst.* **2014**, *55*, 140–147.
13. Gütlein, M.; Frank, E.; Hall, M.; Karwath, A. Large-scale attribute selection using wrappers. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009, Nashville, TN, USA, 30 March–2 April 2009; pp. 332–339.

14. Xu, Y.; Wang, J.; Wei, J. To avoid the pitfall of missing labels in feature selection: A generative model gives the answer. In Proceedings of the AAAI Conference on Artificial Intelligence 2020, New York, NY, USA, 7–12 February 2020; pp. 6534–6541.
15. Chen, W.; Yan, J.; Zhang, B.; Chen, Z.; Yang, Q. Document transformation for multi-label feature selection in text categorization. In Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, Washington, DC, USA, 28–31 October 2007; pp. 451–456.
16. Ma, Z.; Nie, F.; Yang, Y.; Uijlings, J.R.; Sebe, N. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Trans. Multimedia* **2012**, *14*, 1021–1030.
17. Wang, X.; Li, G.Z. Multilabel learning via random label selection for protein subcellular multilocations prediction. *IEEE/ACM Trans. Comput. Biol Bioinform.* **2013**, *10*, 436–446.
18. Zhang, Z.L.; Zhou, Z.H. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1819–1837.
19. Rivolli, A.; J, J.R.; Soares, C.; Pfahringer, B.; de Carvalho, A.C. An empirical analysis of binary transformation strategies and base algorithms for multi-label learning. *Mach. Learn.* **2020**, *9*, 1–55.
20. Zhao, Z.; Wang, L.; Liu, H.; Ye, J. On similarity preserving feature selection. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 619–632.
21. Zhao, J.; Lu, K.; He, X. Locality sensitive semi-supervised feature selection. *Neurocomputing* **2008**, *71*, 1842–1849.
22. Zhang, Y.; Zhou, Z.H. Multi-label dimensionality reduction via dependence maximization. *ACM Trans. Knowl. Discovery Data* **2010**, *4*, 1503–1505.
23. Nie, F.; Xiang, S.; Jia, Y. Trace ratio criterion for feature selection. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, IL, USA, 13–17 July 2008; pp. 671–676.
24. Zhao, Z.; Liu, H. Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th International Conference on Machine Learning, ICML 2007, Corvallis, OR, USA, 20–24 June 2007; pp. 1151–1157.
25. Zhao, Z.; Wang, L.; Liu, H. Efficient spectral feature selection with minimum redundancy. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010; pp. 673–678.
26. Verikas, A.; Bacauskiene, M. Feature selection with neural networks. *Pattern Recog. Lett.* **2002**, *23*, 1323–1335.
27. Arefnezhad, S.; Samiee, S.; Eichberger, A.; Nahvi, A. Driver drowsiness detection based on steering wheel data applying adaptive neuro-fuzzy feature selection. *Sensors* **2019**, *14*, 943.
28. Cateni, S.; Colla, V.; Vannucci, M. A fuzzy system for combining filter features selection methods. *Int. J. Fuzzy Syst.* **2017**, *19*, 1168–1180.
29. Wang, J.; Wei, J.M.; Yang, Z.; Wang, S.Q. Feature selection by maximizing independent classification information. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 828–841.
30. Kong, D.; Ding, C.; Huang, H.; Zhao, H. Multi-label ReliefF and F-statistic feature selections for image annotation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2352–2359.
31. Ji, S.; Ye, J. Linear dimensionality reduction for multi-label classification. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 11–17 July 2009; pp. 1077–1082.
32. Wang, H.; Ding, C.; Huang, H. Multi-label linear discriminant analysis. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 126–139.
33. Jian, L.; Li, J.; Shu, K.; Liu, H. Multi-label informed feature selection. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016; pp. 1627–1633.
34. Huang, J.; Li, G.; Huang, Q.; Wu, X. Joint feature selection and classification for multilabel learning. *IEEE Trans. Cybern.* **2018**, *48*, 876–889.
35. Braytee, A.; Liu, W.; Catchpoole, D.R.; Kennedy, P.J. Multi-label feature selection using correlation information. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 1649–1656.
36. Huang, J.; Li, G.; Huang, Q.; Wu, X. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3309–3323.

37. Ji, S.; Tang, L.; Yu, S.; Ye, J. Extracting shared subspace for multi-label classification. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 381–389.
38. Nie, F.; Huang, H.; Cai, X.; Ding, C.H. Efficient and robust feature selection via joint 2,1-norms minimization. In Proceedings of the 4th Annual Conference on Neural Information Processing Systems 2010, Vancouver, BC, Canada, 6–9 December 2010; pp. 1813–1821.
39. Zhang, M.L.; Wu, L. LIFT: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 107–119.
40. Xiao, Y.H.; Song, H.N. An inexact alternating directions algorithm for constrained total variation regularized compressive sensing problems. *J. Math Imaging Vision* **2012**, *44*, 114–127.
41. Gong, P.; Zhou, J.; Fan, W.; Ye, J. Efficient multi-task feature learning with calibration. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 10–13 August 2014; pp. 761–770.
42. Horn, R.A.; Johnson, C.R. *Matrix Analysis*, 2nd ed.; Cambridge University: Cambridge, UK, 2012.
43. Zhang, M.L.; Zhou, Z.H. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recog.* **2007**, *40*, 2038–2048.
44. Kingma, D.K.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.