Article:

Bai, Lu, Cui, Lixin, Wang, Yue et al. (2 more authors) (2021) Fused Lasso for Feature Selection using Structural Information. Pattern Recognition. 108058. ISSN: 0031-3203

https://doi.org/10.1016/j.patcog.2021.108058

Fused Lasso for Feature Selection using Structural Information

Lixin Cui[1], Lu Bai[1]*, Yue Wang[1], Philip S.Yu[2], Edwin R. Hancock[3]

[1] *Central University of Finance and Economics, Beijing, China.*
[2] *Department of Computer Science, University of Illinois at Chicago, Chicago, USA.*
[3] *University of York, York, UK.*

Abstract

Most state-of-the-art feature selection methods tend to overlook the structural relationship between a pair of samples associated with each feature dimension, which may encapsulate useful information for refining the performance of feature selection. Moreover, they usually consider candidate feature relevancy equivalent to selected feature relevancy, and therefore, some less relevant features may be misinterpreted as salient features. To overcome these issues, we propose a new feature selection method based on graph-based feature representations and the Fused Lasso framework in this paper. Unlike state-of-the-art feature selection approaches, our method has two main advantages. First, it can accommodate structural relationship between a pair of samples through a graph-based feature representation. Second, our method can enhance the trade-off between the relevancy of each individual feature on the one hand and its redundancy between pairwise features on the other. This is achieved through the use of a Fused Lasso framework applied to features reordered on the basis of their relevance with respect to the target feature. To effectively solve the optimization problem, an iterative algorithm is developed to identify the most discriminative features. Experiments demonstrate that our proposed approach can outperform its competitors on benchmark datasets.

Keywords: Feature Selection; Structural Relationship; Fused Lasso

*Correspondence author: Lu Bai, email: bailucs@cufe.edu.cn.

1. Introduction

High-dimensional data are ubiquitous in many data mining and pattern recognition applications [1]. Such data poses significant challenges for classifications, since they not only demand expensive computational complexity but also degrade the generalization ability of the learning algorithm [2]. To tackle this issue, a variety of feature selection methods have been proposed [3]. By eliminating irrelevant and redundant features, feature selection directly chooses a subset of the most salient features from the original feature space so that the classification accuracy and interpretability of the learning algorithm can be improved [4]. In general, feature selection can be partitioned into a) filter methods, b) wrapper methods, and c) embedded methods [3], according to the way of using various learning algorithms in the feature subset selection process. Among these, filters are usually preferred in many real-world applications due to their preferable generalization ability and high computation efficiency [4].

In the literature, many efficient filters have been proposed based on various information theoretic criteria used for evaluating the discriminative power of features, such as correlation [5], mutual information (MI) [6], etc. Among these, MI measures are considered to be most effective as they are able to measure the nonlinear relationships between features and targets [6]. Existing MI methods mostly concentrate on maximizing dependency and relevancy or minimizing redundancy. Representative examples include 1) the mutual information-based feature selection (MIFS) [7], 2) the maximum-relevance minimum-redundancy criterion (MRMR) [8], and 3) the joint mutual information maximisation criterion (JMIM) [9], etc.

Although efficient, most existing information theoretic feature selection methods often utilize measures derived from the statistical characteristics of feature vectors to evaluate their goodness. This may lead to suboptimal solutions for feature selection because the structural relationship between a pair of samples associated with each feature dimension is often neglected by representing features as vectors. However, in many real world applications, structural relationship between pairwise samples associated with each feature dimension may contain useful information that is significant for classification. As an illustrative example (see Figure 1), for three glass-beads denoted as M1

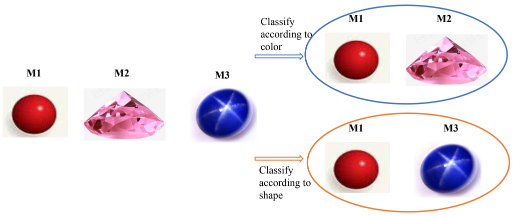


Figure 1: Illustrative Example on Sample Relationship.

(Sphere, Red Colored), M2 (Triangle, Pink Colored), and M3 (Ellipsoid, Blue Colored), the relationships between pairwise samples (M1,M2), (M1,M3), and (M2,M3) are different associated with color. If we classify them according to this feature, M1 and M2 tend to be in the same class. However, if we classify these glass-beads associated with shape, M1 and M3 tend to be in the same class. We can employ any vectorial kernel to compute the kernel-based similarity between these samples and use the C-SVM associated with the kernel matrix for classification. Clearly, the selected features significantly influence the kernel computation between pairs of samples together with the final classification performance. This is because the kernel-based correlation between sample pairs should be greater if the samples are from the same class, and lower if they are from different classes. We observe that the sample relationships of the target feature also satisfy this condition. Thus good fearure selections result if the kernel-based feature graph for a sampled feature is similar to that of its target feature. This indicates that sample relationships associated with the various feature dimensions are important to evaluate the effectiveness of features, and could moreover benefit the classification performance obtained. Therefore, the structural relationships between sample pairs associated with each feature dimension should be incorporated into feature selection.

More specifically, let $\mathbf{f_i} = (f_{i1}, \ldots, f_{ia}, \ldots, f_{ib}, \ldots, f_{iM})^T$ be the vectorial representation for each feature from a dataset consisting of N features and M samples. Traditional information theoretic methods often select the most discriminative features by calculating various criteria such as mutual information between these feature vectors, thus cannot incorporate the relationship between a pair of samples f_{ia} and f_{ib} associated with $\mathbf{f_i}$ into feature selection and may lead to significant information loss.

In feature selection, the appealing characteristics of graph representations have facilitated the development of some pioneering works to tackle this issue. For instance, Zhang et al. [10] have developed a high-order covariate interacted Lasso for feature selection. Specifically, a feature hypergraph is first constructed to characterize the high-order relations among features, where each vertex represents a feature and each hyperedge associated with a weight representing the high-order interaction information among features connected by that hyperedge. A multidimensional interaction information measure is proposed to evaluate the significance of different feature combinations. However, this method cannot incorporate relationship between a pair of samples associated with each feature dimension, thus may lead to significant information loss.

To solve this problem, Cui et al. [11] have introduced a novel feature selection approach based on graph-based feature representations to incorporate relationship between pairwise samples associated with each feature dimension. Specifically, a set of feature graphs is first constructed to incorporate pairwise relationship between samples, where each graph represents each feature. For each feature graph, each vertex denotes a sample, and the edge between a pair of vertices denotes the structural relationship between pairwise samples associated with each feature. With these feature graphs on hand, a novel information theoretic criterion is proposed to evaluate the joint relevancy of different pairwise features. This criterion is utilized to derive an interaction matrix which is further combined with an elastic net model for feature selection.

Although more efficient than [10], the method in [11] suffers from the problem of ignoring some good features with high individual relevancy in relation to the target. For a pair of features which are highly similar and are both relevant to the target, the method in [11] might ignore this pair of features due to the high redundancy degree between them. For instance, given four features with relevancy values 1.0, 0.98, 0.96, and 0.2, the higher the relevancy value, the more discriminative the corresponding feature. For this example, the first feature is correlated with the second and the third features. However, the combination of feature 2 and feature 3 contains more useful information than the first feature, and therefore, it is better to choose feature 2 and feature 3. However, the method in [11] will discard these features because they have high redundancy with the first feature.

In addition, although the Elastic Net regularizer described in [11] can both ensure

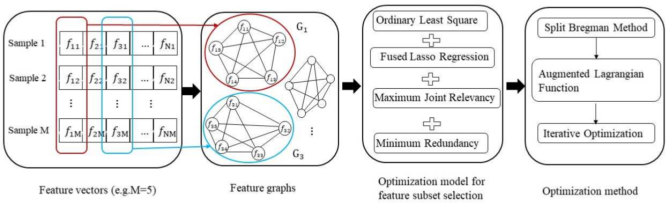


Figure 2: Framework of the proposed feature selection method.

sparsity and promote grouping in the selected features, its performance is less effective when there is an order relation, such as spatial or temporal orders [12] among the features. Tibshirani et al. [13] claimed that features should be ordered while performing feature selection and proposed the Fused Lasso model, which has been shown effective for feature dimension reduction [14]. Specifically, by penalizing the L1-norm of both the coefficients and their successive differences, Fused Lasso can encourage sparsity of both the coefficients and their successive differences, i.e., features that are closely ranked by the order relation will be treated collectively. This indicates that although both Fused Lasso and the Elastic Net can promote a grouping of the selected features, the Elastic Net selects features that are highly correlated as a group while Fused Lasso is more effective for features that can be ordered in a meaningful way.

The purpose of this work is to resolve the aforementioned issues by proposing a novel Fused Lasso feature selection approach using graph-based feature representations. Specifically, our idea is based on converting the original vectorial features into structure-based feature graph representations to incorporate structural relationship between samples, and defining a new structural interaction measure to compute the joint significance of pairwise feature combinations in relation to the target feature graph. With this measure on hand, we obtain a structurally interacting matrix where each element denotes the proposed structural interaction measure. This matrix is used as a regularizer on the feature weights which can simultaneously maximize the joint relevancy and minimize the redundancy of the selected features. Finally, we formulate the corresponding feature subset selection problem into a least square regression model associated with the proposed structurally interacting regularizer and a Fused Lasso regularization term to exploit the ordering effect of regression coefficients which are ranked in terms of relevancy. To effectively solve the corresponding feature subset selection problem, an iterative algorithm is developed to identify the most discriminative features. The framework of the proposed feature selection approach is presented in Figure 2 and the major contributions of this work are highlighted as follows.

First, unlike [11], we illustrate how to convert each original vectorial feature into a structure-based feature graph representation, to encapsulate the structural relationship between a pair of samples. Specifically, a new kernel-based similarity measure associated with the original Euclidean distance is proposed to construct the (target) feature graph structures. Furthermore, a new structural interaction measure associated with the feature graph representations is developed to simultaneously maximize joint relevancy of different pairwise feature combinations in relation to the target feature graphs and minimize redundancy among selected features.

Second, with the proposed structural interaction measure on hand, we compute an interaction matrix to characterize the structural informative relationship between pairwise feature combinations in relation to the target feature graph. Moreover, we formulate the corresponding feature subset selection problem as a least square regression problem associated with a Fused Lasso regularizer. The reasons for using Fused Lasso are as follows. a) When the number of features is larger than the sample size, the maximum number of features selected by Lasso cannot exceed the number of samples. Although Elastic Net can achieve better performance than Lasso, it is less efficient than Fused Lasso when there is an ordering relation of features. b) When there is an ordering relationship for the regression or classification coefficients, Fused Lasso exploits this ordering by explicitly regularizing the differences between neighboring coefficients through an L1-norm regularizer. Thus, it can ensure sparsity not only in the coefficients but also in the differences between neighboring coefficients. That is, for features reordered based upon their individual feature relevancy, Fused Lasso selects several consecutive features which are of high relevancy to the target, and thus enhances the trade-off between relevancy of each individual feature and the redundancy of pairs of features.

Third, because of nonseparability and nonsmoothness of the Fused Lasso regularization term in the objective function, solving the feature selection problem is computationally demanding and difficult. Therefore, an efficient iterative algorithm is developed to locate the optimal solutions to the proposed feature selection problem. The experiments verify the effectiveness of the proposed feature selection approach.

The rest of this paper is organized as follows. In Section 2, we briefly describe the related works. In Section 3, we introduce some important concepts used in the proposed method and present the proposed feature selection method. In Section 4, we exploit an iterative optimization algorithm for solving the proposed model and provide some theoretical analysis on its convergence and computational complexity. In Section 5, we report the experimental results. Finally, In Section 6, we summarize the present study and draw some conclusions.

2. Related Work

Feature selection methods have been extensively investigated in statistical pattern recognition, data mining and machine learning and there have been a number of attempts to review the feature selection methods. For instance, Vergara et al. [15] presents a review of mutual information based feature selection methods. In addition, a comprehensive survey of feature selection algorithms including filter, wrapper, and embedded methods can be found in [3]. In the following, we briefly review state-of-the-art graph-based feature selection methods and regularization-based feature selection methods, which are most relevant to the proposed approach.

2.1. Feature Selection Methods Based on Graphs

Recently, the graph-based approaches, such as semi-supervised learning [16], complex networks [17], and deep learning methods [18], have played a significant role in machine learning due to their capability of encoding the similarity relationships among data. In feature selection, there are mainly two ways of using graphs to model features. The first category represents the feature space as a graph-based representation, and the underlying manifold structure and the relationships between feature vectors can be reflected by a universal and flexible framework. The best known methods are the Fisher score [19] and Laplacian score [20], both of which belong to the graph-based feature selection framework. Many effective graph-based feature selection methods have been proposed within this framework. For instance, Mandal and Mukhopadhyay [21] developed a graph-theoretic approach for selecting non-redundant features without using the class labels of data. In this method, the entire feature space was first converted into a weighted undirected complete graph where the nodes represent the features and the edge weights represent the dissimilarities between the features. Moradi and Rostami [22] proposed a novel graph-theoretic approach for unsupervised feature selection, in which the entire feature set is represented as a weighted graph and these features are further divided into several clusters using a community detection algorithm. Finally, a new iterative search strategy is developed to locate the informative features. However, these methods cannot incorporate the high-order correlation of features into the feature selection process, thus leading to suboptimal solutions.

Zhang and Hancock [23] proposed a hypergraph based information theoretic feature selection approach that overcomes this shortcoming. In their approach, the feature space is characterized as a hypergraph with each node representing a feature and each edge weight corresponding to the multidimensional interaction relationship between the features connected by the edge. This is followed by a hypergraph clustering algorithm which is applied to the hypergraph to identify the most discriminative feature subset. Although much improvement has been achieved, these methods are all based on representing the feature space as a graph, i.e., utilizing the graph-based structure to model the relationship between feature vectors. The structural relationship of a pair of samples in each feature dimension, which encapsulates useful structural information for refining the performance of feature selection, is neglected.

This oversight has lead to some significant advances in the modelling of features as graphs. For instance, in [24], a novel feature selection approach using graph-based features is developed, where each vectorial feature is transformed into a graph-based feature which incorporates the relationship between a pair of samples. More specifically, each vertex denotes a sample, and each edge between two vertices represent the correlation between a pair of samples in the corresponding feature dimension. Then, the most salient vectorial features can be selected by evaluating the graph-based features that are most relevant to the target feature graph, through the Jensen-Shannon

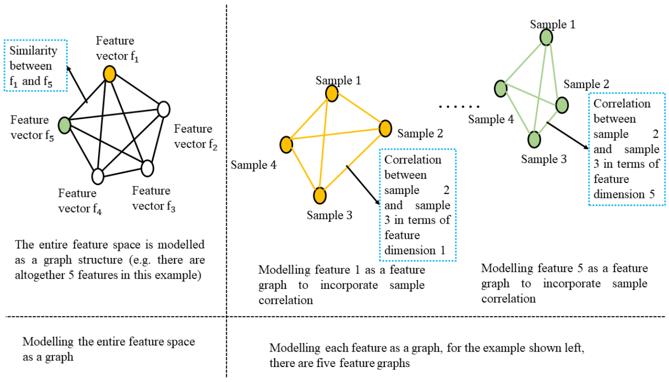

Similarity between f_1 and f_5
Feature vector f_1
Feature vector f_5
Feature vector f_2
Feature vector f_4
Feature vector f_3
The entire feature space is modelled as a graph structure (e.g. there are altogether 5 features in this example)
Sample 1
Sample 2
Sample 4
Sample 3
Correlation between sample 2 and sample 3 in terms of feature dimension 1
Modelling feature 1 as a feature graph to incorporate sample correlation
......
Sample 1
Sample 2
Sample 4
Sample 3
Correlation between sample 2 and sample 3 in terms of feature dimension 5
Modelling feature 5 as a feature graph to incorporate sample correlation
Modelling the entire feature space as a graph
Modelling each feature as a graph, for the example shown left, there are five feature graphs

Figure 3: Differences between feature selection methods based on graphs

divergence between the graphs. However, this approach cannot locate the most discriminative features adaptively, therefore, to overcome this drawback, in [25], a novel information theoretic feature selection approach which can encapsulate the correlation between pairwise samples in each feature dimension is developed. This method can automatically identify the subset containing the most discriminative and less redundant features by solving a quadratic programming problem. These works are quite different from those works which use one graph to represent the entire feature space. The differences of these graph-based feature selection methods are illustrated in Figure 3.

2.2. Regularization-based Feature Selection Methods

In the literature, many effective regularization-based feature selection methods, in particular, the Lasso-type methods such as Lasso, Elastic Net, Group Lasso, etc., have been proposed. Table 1 presents the mathematical formulations of the existing Lasso-type feature selection methods. The methods in the table are described as follows.

Suppose we have a set of training data $\{(\mathbf{x_i}, y_i), i = 1, ..., n\}$ with n observations (samples) and p dimensional features, which is used to estimate the regression coefficients β. Each $\mathbf{x_i} = (x_1^i, x_2^i, ..., x_p^i)^T \in R^{p\times 1}$ is a vector of predictors and $y_i \in R$ is its corresponding response for the i-th sample. We represent the data using the matrix form $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdot\cdot\cdot, \mathbf{x_n}] \in \Re^{p\times n}$ and also represent the response vector as

Table 1: Formulation of existing Lasso-type feature selection methods

Method	Mathematical Formulation
Lasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda\|\beta\|_1$
Elastic Net	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$
Fused lasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\sum_{i=2}^{p}\|\beta_i - \beta_{i-1}\|$
Group Lasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda_1\sum_{g=1}^{G}\|\beta_{I_G}\|_2^2$
ccLasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda\sum_{j=1}^{p}\mu_j\|\beta_j\|, \mu_j = (1 - \|\rho(y, a_j)\|)^2$
unLasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\beta^T\mathbf{C}\beta$
InLasso	$\beta^* = \arg\min_\beta \|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda_1\|\beta\|_1 - \lambda_2\beta^T\mathbf{S}\beta$
InElasticnet	$\beta^* = \arg\min_\beta \frac{1}{2}\|\mathbf{y}^T - \beta^T\mathbf{X}\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 - \lambda_3\beta^T\mathbf{W}\beta$

$\mathbf{y} = (y_1, y_2, ..., y_n)^T \in \Re^n$. With this representation to hand, the linear regression model, for instance, can be written as

$$\min_{\beta\in\Re^p}\sum_{i=1}^{n}\|y_i - \sum_{j=1}^{p}\beta_j x_j^i\|_2^2 = \min_{\beta\in\Re^p}\|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2$$
$$s.t. \sum_{i=1}^{n}\|\beta\|_0 = k, \tag{1}$$

where $\mathbf{y} \in \Re^{n\times 1}$ represents the label vector, $\mathbf{X} \in \Re^{p\times n}$ denotes the training dataset, k denotes a predetermined number of the selected features. To estimate the regression coefficients β and fit the above linear regression model, the ordinary least square (OLS) approach is used. OLS selects the coefficients $\beta = (\beta_1, ..., \beta_p)^T$ by minimizing the residual sum of squares shown in Eq.(1). However, the minimisation of Eq.(1) is an NP-hard optimization problem which is difficult to be solved exactly. Therefore, we can relax the constraint equation in problem Eq.(1) by imposing a positive regularization parameter λ and adding it to the objective function, that is

$$\min_{\beta\in\Re^p}\|\mathbf{y}^T - \beta^t\mathbf{X}\|_2^2 + \lambda\|\beta\|_0. \tag{2}$$

However, solving Eq.(2) is still very challenging. Hence, an alternative formulation using l_1-norm regularization instead of the l_0-norm has been proposed for practical purposes. This corresponds to the regularized counterpart of the Lasso problem in statistical learning [26]. Lasso imposes an l_1 constraint on the regression weights, such that some regression weights in the regression model will shrink to zero. Therefore, it automatically chooses a set of the discriminative variables. The sparsity is controlled by the tuning parameter $\lambda \geq 0$. As λ becomes larger, the larger the number of zero elements in the vector β, i.e., the features are selected more sparsely.

However, Lasso is based on the independence assumption on the input variables, though for a great variety of real world applications features are often correlated. Hence, when highly correlated features exist, Lasso tends to choose only one of these features, leading to suboptimal performance. To overcome this drawback, the Elastic Net [27] uses an additional l_2 regularization term into the Lasso objective function for selecting groups of correlated features. Thus the Elastic Net can be seen as a linear combination of the Lasso and Ridge penalty. It thus enjoys a similar sparsity of representation with Lasso and also promotes a grouping effect on the selected features. This leads to a more appropriate feature selection and predictive performance.

Given feature grouping information, the Group Lasso [28] attempts to conduct feature selection on groups of variables. It performs Lasso at an inter-group level, where features from different groups compete to survive, and an entire group of predictors may be dropped out of the model simultaneously. For a group of features with high pairwise correlations, the Lasso tends to choose only one feature from the group and is not sensitive to the features selected. In contrast, the Group Lasso determines whether this group of features is discriminative. Nonetheless, the Group Lasso model requires a non-overlapping group structure, which restricts its practical applicability [10].

When there is some ordering relationship among the coefficients and all the coefficients are closely related to their neighbors, Fused Lasso [13] enforces sparsity in both the coefficients and their successive differences. It thus yields a solution with sparsity in both the coefficients and their successive differences. Fused Lasso is especially useful when the number of features p is much greater than the sample size n. However, comparing to the alternative models, Fused Lasso is more difficult to solve due to the nonseparability and nonsmoothness of the regularization term in the objective function.

The above review indicates that traditional sparse learning feature selection methods are based on the assumption that there is conditional independence among the variables, and they aim to perform regression individually for each response vector. Therefore, they fail to incorporate correlation between the variables and response as well as the variable correlation into the feature selection process. Recently, some methods have been proposed to overcome this drawback. For example, Chen et al. [29] developed an uncorrelated Lasso (unLasso) to perform variable de-correlation and feature selection simultaneously, such that the features selected are uncorrelated as much as possible, leading to less redundancy. Moreover, Jiang et al. [30] developed a covariate-correlated Lasso (ccLasso) which identifies the covariates that correlates more strongly with the target. As a result, the features selected are highly relevant to the target, resulting in high relevance. Furthermore, to incorporate high-order feature interactions into a Lasso regression model, Zhang et al. [10] developed a high-order covariate interacted lasso (InLasso) feature selection method. By conducting a feature hypergraph to model the high-order interactions among features and utilizing the feature hypergraph as a regularizer on the feature weights, InLasso automatically determines whether a feature is redundant or interactive depending on a neighborhood dependency criterion.

Although the above methods have improved the performance of feature selection to some extent, the selected features may not be optimal. This is because none of the above works can incorporate the structural relationship of pairwise samples in each feature dimension into the feature selection process. Intuitively, such structural information is one type of prior information that can benefit the feature selection problem. To resolve this problem, Cui et al. [11] have developed a novel graph-based feature selection method. They commence by transforming each vectorial feature into a graph-based representation which incorporates structural relationship between a pair of samples. A new structural interaction measure is developed to quantify the joint relevancy of different pairwise feature combinations in relation to target graph features. Then a new optimization model associated with a least square error, an elastic net regularizer, and the proposed interaction measure is formulated to select the discriminative feature subsets. Although efficient, the proposed method has the following issues. First, it adopts the Euclidean distance as the measure to construct both the feature graph $\mathbf{G_i}(V_i, E_i)$ and the target feature graph $\hat{\mathbf{G}}_\mathbf{i}(\hat{V}_i, \hat{E}_i)$ to compute the relationship between a pair of feature samples. However, the characteristics of these graph structures may be overemphasized and dominated by the large distance value. In addition, for a pair of features which are highly similar and are both relevant to the target, [11] might ignore this pair of features due to the high redundancy between them.

In this paper, we propose a new feature selection method to overcome these issues.

We commence by introducing some important preliminary concepts as below.

3. Preliminary Concepts

In this section, we first illustrate the construction of the feature graph which incorporates structural information of pairwise feature samples. We then review the preliminaries of Jensen-Shannon divergence for multiple probability distributions, which is utilized to compute the similarity between feature graph structures.

3.1. Kernel-based Feature Graph Modelling

In this subsection, we illustrate how to convert the original vectorial features into structure-based feature graphs, in terms of a kernel-based similarity measure. The reason of representing each original feature as a graph structure is that graph-based representation can capture richer global topological information than vectors. Thus, the pairwise sample relationships of each original feature vector can be incorporated into the selection of the most discriminative features to reduce information loss.

Let $\mathcal{X} = \{\mathbf{f_1}, \ldots, \mathbf{f_i}, \ldots, \mathbf{f_N}\} \in R^{M \times N}$ be a dataset of N features and M samples. We transform each original vectorial feature $\mathbf{f_i} = (f_{i1}, \ldots, f_{ia}, \ldots, f_{ib}, \ldots, f_{iM})^T$ into a feature graph structure $\mathbf{G_i}(V_i, E_i)$, where each vertex $v_{ia} \in V_i$ represents the a-th sample f_{ia} and each weighted edge $(v_{ia}, v_{ib}) \in E_i$ represents the relationship between the a-th and b-th samples. Moreover, we also need to construct a graph structure for the target feature $\mathbf{Y}$. For classification problems, $\mathbf{Y}$ are the class labels and usually take the discrete class values $c \in \{1, 2, \ldots, C\}$. For such case, we first compute the continuous value based target feature for each feature $\mathbf{f_i}$ as $\hat{\mathbf{f}}_\mathbf{i} = (\hat{f}_{i1}, \ldots, \hat{f}_{ia}, \ldots, \hat{f}_{ib}, \ldots, \hat{f}_{iM})^T$, where each element $\hat{f}_{ia}$ corresponds to the a-th sample. When the element f_{ia} of $\mathbf{f_i}$ belongs to the c-th class, the value of $\hat{f}_{ia}$ is the mean value μ_{ia} of all samples in $\mathbf{f_i}$ from the same class c. Similar to the process of converting each original feature $\mathbf{f_i}$ into the feature graph, we construct the resulting target feature graph representation for each feature $\mathbf{f_i}$ associated with its continuous value based target feature $\hat{\mathbf{f}}_\mathbf{i}$ as $\hat{\mathbf{G}}_\mathbf{i}(\hat{V}_i, \hat{E}_i)$, where each vertex $\hat{v}_{ia}$ represents the a-th sample of $\hat{\mathbf{f}}_\mathbf{i}$ (i.e., the a-th sample of $\mathbf{Y}$ in terms of $\hat{\mathbf{f}}_\mathbf{i}$), and $(v_{ia}, v_{ib}) \in E_i$ represents the relationship between the a-th and b-th samples of $\mathbf{f_i}$ (i.e., the structural relationship between the a-th and b-th samples of $\mathbf{Y}$ in terms of $\hat{\mathbf{f}}_\mathbf{i}$). To compute the relationship between pairwise samples, [11] employed the Euclidean distance as the measure to construct both the feature graph $\mathbf{G_i}(V_i, E_i)$ and the target feature graph $\hat{\mathbf{G}}_\mathbf{i}(\hat{V}_i, \hat{E}_i)$. However, the characteristics of these graph structures may be overemphasized and dominated by the large distance value.

To overcome the aforementioned problem, we further propose a new kernel-based similarity measure associated with the original Euclidean distance to construct the (target) feature graph structures. Specifically, for the feature graph $\mathbf{G_i}(V_i, E_i)$ of $\mathbf{f_i}$ and its associated Euclidean distance based adjacency matrix A, each row (column) of A can be seen as the distance based embedding vector for each sample of $\mathbf{f_i}$. Assume $A_{a,:}$ and $A_{b,:}$ denote the embedding vectors of the a-th and b-th samples respectively. The relationship between these two samples can be computed as their normalized kernel value associated with dot product

$$K_{a,b} = \frac{\langle A_{a,:}, A_{b,:}\rangle}{\sqrt{\langle A_{a,:}, A_{a,:}\rangle\langle A_{b,:}, A_{b,:}\rangle}}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is the dot product. We utilize the kernel matrix to replace the original Euclidean distance matrix as the adjacency matrix of $\mathbf{G_i}(V_i, E_i)$, and the relationships between the samples of $\mathbf{f_i}$ are all bounded between 0 and 1. For the target feature graph $\hat{\mathbf{G}}_\mathbf{i}(\hat{V}_i, \hat{E}_i)$, we also compute its adjacency matrix using the same procedure. The kernel-based similarity measure not only overcomes the shortcoming of graph characteristics domination by the large Euclidean distance value between pairwise feature samples, but also encapsulates high-order relationship between feature samples. This is because the kernel-based relationship between each pair of samples associated with their distance based embedding vector encapsulates the distance information between each feature sample and the remaining feature samples. Finally, the kernel-based relationship can also represent the original vectorial features in a high-dimensional Hilbert space, and thus reflect richer structural characteristics.

3.2. The JSD for Multiple Probability Distributions

In Statistics and Information Theory, an extensively used measure of dissimilarity between probability distributions is the Jensen Shannon divergence (JSD) [31]. JSD has been successful in a wide range of applications, including analysis of symbolic sequences and segmentation of digital images. In [32], the JSD has been adopted to measure similarity between graphs associated with their probability distributions. Moreover, [11] have utilized the JSD to compute the similarity between an individual feature graph in relation to its target feature graph. Unlike the previous works that focus on the JSD measure between pairwise graph structures, our major concern is the similarity between multiple graphs. Specifically, the JSD measure can be used to compare n ($n \geq 2$) probability distributions,

$$D_{\mathrm{JS}}(\mathcal{P}_1, \cdots, \mathcal{P}_n) = H_S\Big(\sum_{i=1}^{n} \pi_i \mathcal{P}_i\Big) - \sum_{i=1}^{n} \pi_i H_S(\mathcal{P}_i) \tag{4}$$

where $\pi_i \geq 0$ is the corresponding weight for the probability distribution $\mathcal{P}_i$ and $\sum_{i=1}^{n} \pi_i = 1$. In addition, H_S denotes the Shannon entropy of a probability distribution. In this work, we set each $\pi_i = \frac{1}{n}$. Since we aim to calculate the joint relevancy between features in terms of similarity measures between graph-based feature representations, we utilize the negative exponential of D_{JS} to calculate the similarity I_S between the multiple n ($n \geq 2$) probability distributions, i.e.,

$$I_S(\mathcal{P}_1, \cdots, \mathcal{P}_n) = \exp\{-D_{\mathrm{JS}}(\mathcal{P}_1, \cdots, \mathcal{P}_n)\}. \tag{5}$$

4. The Proposed Fused Lasso Feature Selection Using Structural Information

In this section, we introduce the proposed approach. We commence by defining a new structural interaction measure to compute the joint relevancy between features. Moreover, we present the mathematical formulation for the feature subset selection problem and exploits an iterative optimization algorithm to solve it.

4.1. The Proposed Structural Interaction Measure

We propose the following structural interaction measure for evaluating the joint relevancy of different pairwise feature combinations in relation to the target features. For the set of N features $\mathbf{f_1}, \ldots, \mathbf{f_i}, \ldots, \mathbf{f_j}, \ldots, \mathbf{f_N}$ defined earlier and the associated discrete target feature $\mathbf{Y}$ taking the discrete values $c \in \{1, 2, \ldots, C\}$, we calculate the joint relevance degree of the feature pair $\{\mathbf{f_i}, \mathbf{f_j}\}$ in relation to the target feature $\mathbf{Y}$ as

$$U_{\mathbf{f_i},\mathbf{f_j}} = \frac{I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{i}) + I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{j})}{I_S(\mathbf{G_i}, \mathbf{G_j})} \tag{6}$$

where $\mathbf{G_i}$ is the feature graph of each original feature $\mathbf{f_i}$, $\hat{\mathbf{G}}_\mathbf{i}$ is the target feature graph of $\mathbf{Y}$ in terms of $\mathbf{f_i}$. $I_S(\mathbf{G_i}, \mathbf{G_j})$ and $I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{i})$ are the JSD-based information theoretic similarity measures calculated by Eq.(5) for $n = 2$ and $n = 3$, respectively. The above interaction measure is composed of two terms. The first term $I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{i}) + I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{j})$ measures the relevancy of pairwise features $\mathbf{f_i}$ and $\mathbf{f_j}$ in relation to the target feature $\mathbf{Y}$. The second part $I_S(\mathbf{G_i}, \mathbf{G_j})$ evaluates the redundancy between the feature pair $\{\mathbf{f_i}, \mathbf{f_j}\}$. As a result, the proposed structural interaction measure $U_{\mathbf{f_i},\mathbf{f_j}}$ is large if and only if $I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{i}) + I_S(\mathbf{G_i}, \mathbf{G_j}; \hat{\mathbf{G}}_\mathbf{j})$ is large and $I_S(\mathbf{G_i}, \mathbf{G_j})$ is small. This indicates that the pairwise features $(\mathbf{f_i}, \mathbf{f_j})$ are informative and less redundant.

Although the proposed structural interaction measure as well as that proposed by [11] are both related to the JSD measure, the proposed measure differs from [11] in that our method focuses on the JSD measure between multiple probability distributions rather than only two probability distributions to compute the feature relevance. Therefore, the proposed structural interaction measure can compute the joint relevancy of a pair of feature combinations in relation to the target. By contrast, the measure proposed by [11] is based upon the relevance degree of each individual feature in relation to the target feature graph, which may result in the selection of less relevant features.

Moreover, based upon the graph-based feature representations, we obtain a structural information matrix $\mathbf{U}$, where each entry $U_{i,j} \in \mathbf{U}$ corresponds to the structural interaction measure between a pair of features $\{\mathbf{f_i}, \mathbf{f_j}\}$ based on Eq.(6). Given the structural information matrix $\mathbf{U}$ and the N-dimensional feature coefficient vector β, where β_i corresponds to the coefficient of the i-th feature, one can locate the most discriminative feature subset by solving the optimization problem below

$$\max f(\beta) = \max_{\beta \in \Re^N} \beta^T \mathbf{U} \beta, \tag{7}$$

where $\beta \geq 0$.

4.2. Mathematical Formulation

Our feature selection approach aims to capture structural information between pairwise features and encourage the selected features to be jointly more relevant with the target while maintaining little redundancy. In addition, it should simultaneously promote a sparse solution both in the coefficients and their successive differences. Therefore, we unify the minimization problem of Fused Lasso and Eq.(7) and propose the fused lasso for feature selection using structural information(InFusedLasso), which is mathematically formulated as

$$\min_{\beta \in \Re^N} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\mathbf{C}\beta\|_1 - \lambda_3\beta^T\mathbf{U}\beta, \tag{8}$$

where $\mathbf{y} \in \Re^{n\times 1}$ represents the label vector, $\mathbf{X} \in \Re^{p\times n}$ denotes the training dataset, λ_1 and λ_2 are the tuning parameters for the Fused Lasso model, and λ_3 is the corresponding tuning parameter of the structural interaction matrix $\mathbf{U}$. The first term in the above objective function is the ordinary least square error term which utilizes information from the original feature space. The second regularization term with parameter λ_1 encourages the sparsity of β as in Lasso and the third regularization term with parameter λ_2 shrinks the differences between successive features specified in matrix $\mathbf{C}$ toward zero. Same as in standard Fused Lasso [33], $\mathbf{C}$ is a $(N-1)\times N$ matrix with zero entries everywhere except 1 in the diagonal and -1 in the superdiagonal. Moreover, the fourth term encourages the selected features to be jointly more relevant with the target while maintaining less redundancy among them. To solve the proposed model (8), it is of great necessity to develop an efficient and effective algorithm to locate the optimal solutions, i.e., β^*. A feature $\mathbf{f_i}$ belongs to the optimal feature subset if and only if $\beta_i^* > 0$. Accordingly, the number of optimal features can be recovered based on the number of positive components of β^*.

4.3. Optimization Algorithm

To effectively resolve model (8), we develop an optimization algorithm based upon the split Bregman iteration approach [33]. We commence by reformulating the unconstrained problem (8) into an equivalent constrained problem shown below

$$\begin{aligned} &\min_{\beta \in \Re^N} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|p\|_1 + \lambda_2 \|q\|_1 - \lambda_3 \beta^T \mathbf{U} \beta \\ &\text{s.t.} \quad p = \beta, \\ &\qquad\ \ q = \mathbf{C}\beta. \end{aligned} \tag{9}$$

To solve the problem, we derive the split Bregman method for the proposed optimization model (9) using the augmented Lagrangian method [34]. To be specific, the corresponding Lagrangian function of (9) is

$$\begin{aligned} \widetilde{L}(\beta, p, q, u, v) =& \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 - \lambda_3 \beta^T \mathbf{U} \beta + \lambda_1 \|p\|_1 \\ & + \lambda_2 \|q\|_1 + \langle u, \beta - p \rangle + \langle v, \mathbf{C}\beta - q \rangle, \end{aligned} \tag{10}$$

where $u \in \Re^N$ and $v \in \Re^{N-1}$ are the dual variables corresponding to the linear constraints $p = \beta$ and $q = \mathbf{C}\beta$, respectively. Here $\langle \cdot, \cdot \rangle$ denotes the standard inner product in the Euclidean space. By adding two terms $\frac{\mu_1}{2}\|\beta - p\|_2^2$ and $\frac{\mu_2}{2}\|\mathbf{C}\beta - q\|_2^2$ to penalize the violation of linear constraints $p = \beta$ and $q = \mathbf{C}\beta$, one can obtain the augmented Lagrangian function of (10), that is,

$$\begin{aligned} L(\beta, p, q, u, v) =& \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 - \lambda_3 \beta^T \mathbf{U} \beta + \lambda_1 \|p\|_1 \\ & + \lambda_2 \|q\|_1 + \langle u, \beta - p \rangle + \langle v, \mathbf{C}\beta - q \rangle \\ & + \frac{\mu_1}{2}\|\beta - p\|_2^2 + \frac{\mu_2}{2}\|\mathbf{C}\beta - q\|_2^2, \end{aligned} \tag{11}$$

where $\mu_1 > 0$ and $\mu_2 > 0$ are the corresponding parameters. To find a saddle point denoted as $(\beta^*, p^*, q^*, u^*, v^*)$ for the augmented Lagrangian function Ł(β, p, q, u, v), the following inequalities hold

$$L(\beta^*, p^*, q^*, u, v) \leq L(\beta^*, p^*, q^*, u^*, v^*) \leq L(\beta, p, q, u^*, v^*), \tag{12}$$

for all β, p, q, u and v. It is clear that β^* is an optimal solution to (8) if and only if $\beta^*, p^*, q^*, u^*, v^*$ solves this saddle point problem for some p^*,q^*,u^*,and v^* [35].

We solve the above saddle point problem using an iterative algorithm by alternating between the primal and the dual optimization shown below

$$\begin{cases} \text{Primal:}(\beta^{k+1}, p^{k+1}, q^{k+1}) = \arg\min\limits_{\beta,p,q} L(\beta, p, q, u^*, v^*) \\ \text{Dual:}u^{k+1} = u^k + \delta_1(\beta^{k+1} - p^{k+1}) \\ \qquad v^{k+1} = v^k + \delta_2(\mathbf{C}\beta^{k+1} - q^{k+1}), \end{cases}$$

where the first step updates the primal variables based upon the current estimation of u^k and v^k, followed by the second step which updates the dual variables based upon the current estimates of the primal variables. Because the augmented Lagrangian function is linear in both u and v, updating the dual variables is comparatively simple and we adopt a gradient ascent method with step size δ_1 and δ_2. Therefore, the efficiency of the above optimization algorithm depends upon whether the primal problem can be resolved quickly. To facilitate better illustration, denote

$$\begin{aligned} V(\beta) =& \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 - \lambda_3\beta^T\mathbf{U}\beta + \langle u^k, \beta - p^k\rangle + \langle v^k, \mathbf{C}\beta - q^k\rangle \\ &+ \frac{\mu_1}{2}\|\beta - p^k\|_2^2 + \frac{\mu_2}{2}\|\mathbf{C}\beta - q^k\|_2^2. \end{aligned} \tag{13}$$

Because the objective function on minimizing β, i.e.,$V(\beta)$ is differentiable, we can resolve the primal problem by alternatively minimizing β, p, and q as follows,

$$\begin{cases} \beta^{k+1} = \arg\min\limits_{\beta} V(\beta^k) \\ p^{k+1} = \arg\min\limits_{p} \lambda_1\|p\|_1 + \langle u^k, \beta^{k+1} - p\rangle + \frac{\mu_1\|\beta^{k+1} - p\|_2^2}{2} \\ q^{k+1} = \arg\min\limits_{q} \lambda_2\|q\|_1 + \langle v^k, \mathbf{C}\beta^{k+1} - q\rangle + \frac{\mu_2\|\mathbf{C}\beta^{k+1} - q\|_2^2}{2}. \end{cases} \tag{14}$$

Furthermore, since the objective function $V(\beta)$ on minimizing β is quadratic and differentiable, we can obtain the optimal solution of β by setting $\frac{\partial(V(\beta))}{\partial\beta} = 0$, that is,

$$\begin{aligned} &\mathbf{X}^T\mathbf{X}\beta - 2\lambda_3\mathbf{U}\beta + \mu_1\mathbf{I}\beta + \mu_2\mathbf{C}^T\mathbf{C}\beta - \mathbf{X}^Ty - \mu_1p^k \\ &+\mu_1\mu_1^{-1}u^k - \mu_2\mathbf{C}^Tq^k + \mu_2\mathbf{C}^T\mu_2^{-1}v^k = 0, \end{aligned} \tag{15}$$

i.e.,the optimal solution are obtained by solving a set of linear equations as follows

$$\mathbf{D}\beta^{k+1} = \mathbf{X}^Ty + \mu_1(p^k - \mu_1^{-1}u^k) + \mu_2\mathbf{C}^T(q^k - \mu_2^{-1}v^k). \tag{16}$$

Because matrix $\mathbf{D} = \mathbf{X}^T\mathbf{X} - 2\lambda_3\mathbf{U} + \mu_1\mathbf{I} + \mu_2\mathbf{C}^T\mathbf{C}$ is an $N \times N$ matrix, which is independent of the optimization variables. For small N, we can invert $\mathbf{D}$ and store $\mathbf{D}^{-1}$ in the memory, such that the linear equations are resolved with minimum cost. That is, $\beta^{k+1} = \mathbf{D}^{-1}(\mathbf{X}^T y + \mu_1(p^k - \mu_1^{-1}u^k) + \mu_2\mathbf{C}^T(q^k - \mu_2^{-1}v^k))$. However, for large N, we need to numerically solve the linear equations at each iteration by means of the conjugate gradient algorithm.

In addition, the objective functions of the minimization of p and q are quadratic and nondifferentiable terms according to Eq.(14), which are completely separable, therefore we adopt the soft thresholding approach to find the optimal solutions for p and q. Specifically, let t_λ be a soft thresholding operator defined on vector space which satisfies $\Gamma_\lambda(\omega) = [t_\lambda(\omega_1, t_\lambda(\omega_1, ..., ...]^T$, with $t_\lambda(\omega_i) = \text{sgn}(\omega_i)\max\{0, |\omega_i| - \lambda\}$.

Using the soft thresholding operator, the optimal solution of p and q in Eq.(12) can be respectively calculated as

$$p^{k+1} = \Gamma_{\mu_1^{-1}\lambda_1}(\beta^{k+1} + \mu_1^{-1}u^k), \tag{17}$$

and

$$q^{k+1} = \Gamma_{\mu_2^{-1}\lambda_2}(\mathbf{C}\beta^{k+1} + \mu_2^{-1}v^k). \tag{18}$$

Moreover, according to Eq.(13), the optimal solution of u and v can be respectively updated as

$$u^{k+1} = u^k + \delta_1(\beta^{k+1} - p^{k+1}), \tag{19}$$

and

$$v^{k+1} = v^k + \delta_2(\mathbf{C}\beta^{k+1} - q^{k+1}). \tag{20}$$

Overall, we develop Algorithm 1 for locating optimal solutions to the proposed feature selection problem.

4.4. Computational Complexity

Suppose N is the number of features, M is the number of samples, and K is the required number of iterations to converge at optima. For each iteration, the computational complexity for updating β according to Eq.(16) is $O(N^2M)$. Additionally, the computational costs for updating p in Eq.(17) and q in Eq.(18) are both $O(N)$. Moreover, the computational costs for updating u in Eq.(19) and v in Eq.(20) are both $O(N)$.

Algorithm 1 The iterative optimization algorithm for the feature selection problem

Input: $\mathbf{X}, \mathbf{y}, \beta^0, p^0, q^0, u^0$ and v^0.

Output: β^*

```
while not converged do
    Update $\beta^{k+1}$ according to the solution to Eq.(16).
    Update $p^{k+1}$ using Eq.(17).
    Update $q^{k+1}$ using Eq.(18).
    Update $u^{k+1}$ using Eq.(19).
    Update $v^{k+1}$ using Eq.(20).
end while
return solution
```

Therefore, the overall time complexity of the proposed iterative algorithm is calculated as $\max\{O(N^2MK), O(NK)\}$.

4.5. Convergence Proof

In this subsection, we present the convergence property of Algorithm 1.

Theorem 1. Assume there exists at least one solution denoted as β^* of the optimization problem (8). Suppose $V(\beta)$ is convex, $0 < \delta \leq \mu_1$, $0 < \delta_2 \leq \mu_2$, and $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 > 0$. To facilitate the presentation, assume $J(\beta) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 - \lambda_3\beta^T\mathbf{U}\beta$. Then the following property for the split Bregman iteration shown in Algorithm 1 holds, i.e.,

$$\lim_{k\to\infty} J(\beta) + \lambda_1\|\beta^k\|_1 + \lambda_2\|L\beta^k\|_1 = J(\beta^*) + \lambda_1\|\beta^*\|_1 + \lambda_2\|L\beta^*\|_1. \tag{21}$$

Moreover,

$$\lim_{k\to\infty} \|\beta^k - \beta^*\| = 0, \tag{22}$$

whenever problem (8) has a unique solution. Note that the condition for the convergence in **Theorem 1** is quite easy to satisfy. This is because the regularization parameters λ_1 and λ_2 should always be greater than zero. Therefore, as long as $0 < \delta_1 \leq \mu_1$ and $0 < \delta_2 \leq \mu_2$, the algorithm converges. In our implementation, we just choose

Table 2: Statistics of experimental datasets

Dataset	USPS	Pie	YaleB	Lymphoma	Leukemia	RELATHE	BASEHOCK	Isolet1
#Feature	256	1024	1024	4026	7129	4322	4862	617
#Sample	9298	11554	2414	96	73	1427	1993	1560
#Class	10	68	38	9	22	2	2	26
Type	Image	Image	Image	Biomedical	Biomedical	Text	Text	Speech

$\delta_1 = \mu_1$ and $\delta_2 = \mu_2$. We will provide the convergence proof for **Theorem 1** in Appendix, which follows similar ideas from [36].

5. Experiments

In this section, we conduct several experiments on standard machine learning datasets to verify the performance of the proposed Fused Lasso feature selection method, i.e., InFusedLasso(K), where K denotes the kernel-based feature modelling approach. The purposes of the experiments are to: 1) compare the proposed algorithms with several benchmark feature selection methods to demonstrate the performance of our method; 2) compare the results obtained with and without the kernel-based feature modelling approach to show its effectiveness, especially for the data sets associated with features re-ordered according to their individual feature relevance to the target; 3) conduct convergence analysis of the proposed method; and 4) conduct a parameter study in order to choose the optimal parameter settings for the experiments.

More specifically, we use eight standard machine learning datasets abstracted from Biomedical, Speech, Text and Computer Vision databases. The numbers of features vary from 256 to 7129 with a mean of 2908. The dimensionality of half of these datasets exceed 4,000. In addition, the numbers of samples vary from 73 to 11,554. Among them, RELATHE and BASEHOCK are both large in feature dimension and sample size, whereas Lymphoma and Leukemia are typical datasets with high-dimensional features and small sample size. Details of these datasets are presented in Table 2.

5.1. Experimental Settings

To evaluate the performance of our proposed InFusedLasso(K) method and compare it with state-of-the-art feature selection methods in a fair and reasonable way, we set up our experiments as follows. The proposed method is compared with several existing state-of-the-art Lasso-type feature selection methods and one graph-based feature selection method, i.e., InElasticNet [11]. The Lasso-type methods used for comparisons include Lasso [26], Fused Lasso [13], Group Lasso [28], ULasso [29], InLasso [10], and ccLasso [30].

i) Lasso performs feature selection through the l_1-norm, where features corresponding to zero coefficients in the parameter vector will be discarded.

ii) ULasso aims to conduct variable de-correlation and variable selection simultaneously, such that the variables selected are uncorrelated as much as possible.

iii) Fused Lasso encourages sparsity in both the coefficients and their successive differences, which is useful for applications with features ranked in meaningful ways.

iv) Group Lasso can enforce sparsity on features at an inter-group level, where features from different groups compete with each other and will be in and out of the model as a group.

v) Elastic Net linearly combines the l_1 and l_2 regularization terms of the Lasso and Ridge approaches. It ensures democracy among groups of correlated groups and allows selection of the relevant groups while promoting sparse solutions.

vi) InLasso encapsulates high-order feature interactions, which effectively evaluates whether a feature is redundant or interactive based on a neighborhood dependency criterion. It avoids deleting useful features arising in individual feature combinations.

vii) ccLasso applies prior knowledge of variable-response correlation into the Lasso regularized feature selection method, so that the features chosen can be strongly correlated with the response.

viii) InElasticNet is a graph-based feature selection method which incorporates pairwise relationship between samples of each feature dimension.

In order to make the best use of the data and obtain stable results, a 10×10-fold cross-validation strategy is used. Specifically, for each dataset, each feature selection algorithm associated with a C-SVM classifier based on the Linear kernel, the 10-fold cross-validation approach is repeated 10 times. For the 10-fold cross-validation approach, we use nine folders for training and one folder for testing. In the experiments, we vary the number of selected features from zero to the total number of features of

Table 3: The best results of all methods and the corresponding number of selected features.

Datasets	Lasso	ULasso	FusedLasso	GroupLasso	InLasso	**InFusedLasso(D)**	**InFusedLasso(K)**
USPS	86.30(50)	83.25(50)	87.40(50)	83.93(50)	93.94(50)	92.50(50)	**94.16**(50)
Pie	94.48(70)	94.57(70)	86.94(70)	92.35(160)	96.58(70)	96.00(70)	**96.90**(70)
YaleB	46.64(50)	47.43(50)	48.09(50)	45.02(50)	71.20(50)	**95.24**(50)	**96.37**(50)
Lymphoma	91.11(100)	94.44(160)	90.00(120)	91.11(200)	96.00(140)	96.00(140)	**96.55**(120)
Leukemia	82.86(200)	82.86(200)	94.29(140)	91.43(180)	**100.00**(80)	**100.00**(80)	**100.00**(100)
RELATHE	86.00(200)	85.49(200)	85.62(200)	74.33(200)	80.70(180)	**86.83**(180)	**86.59**(200)
BASEHOCK	67.22(140)	67.33(200)	84.62(200)	73.33(200)	86.58(180)	**93.87**(180)	**94.51**(200)
Isolet1	91.67(100)	92.18(100)	88.08(90)	83.53(100)	91.92(100)	91.92(100)	**93.26**(100)

Table 4: InFusedLasso versus InElasticNet.

Datasets	InElasticNet	**InFusedLasso(D)**	**InFusedLasso(K)**
USPS	94.10(50)	92.50(50)	**94.16**(50)
Pie	96.81(70)	96.00(70)	**96.90**(70)
YaleB	94.62(50)	**95.24**(50)	**96.37**(50)
Lymphoma	95.56(160)	**96.00**(140)	**96.55**(120)
Leukemia	**100.00**(80)	**100.00**(80)	**100.00**(100)
RELATHE	83.66(180)	**86.83**(200)	**86.59**(200)
BASEHOCK	92.75(200)	**93.87**(200)	**94.51**(200)
Isolet1	92.23(100)	91.92(100)	**93.26**(100)

the dataset, with a fixed interval, e.g., 5, 10, 20, etc., to investigate the changes in the classification accuracy associated with different number of features [10, 37]. And the performance of various feature selection methods is evaluated in terms of the mean classification accuracies versus different number of selected features.

5.2. Comparison of Classification Accuracy with Standard Feature Selection Methods

The classification performance obtained via the comparative methods are shown in Figure 4. In addition, the best mean classification accuracies of different methods associated with the number of selected features are reported in Table 3.

Figure 4 exhibits the classification accuracies of different algorithms obtained with different number of selected features. From this figure, it can be noticed that when the number of selected features reaches a certain number, the proposed approach can outperform the alternative methods on all the eight datasets, which demonstrates the advantage of the proposed InFusedLasso method. Moreover, Table 3 confirms that the proposed approach can achieve the best classification performance on all the eight datasets. Additionally, the proposed InFusedLasso (K) method can achieve on average 2.96% to 37.54% improvement for all the six baseline methods including Lasso, ULas-

Table 5: InFusedLasso on representative datasets before and after feature pre-ordering.

Dataset	YaleB	Lymphoma	Leukemia	Isolet1
Unordered	94.48(50)	95.55(160)	98.57(200)	93.01(100)
Ordered	**96.37**(50)	**96.55**(120)	**100.00**(100)	**93.26**(100)

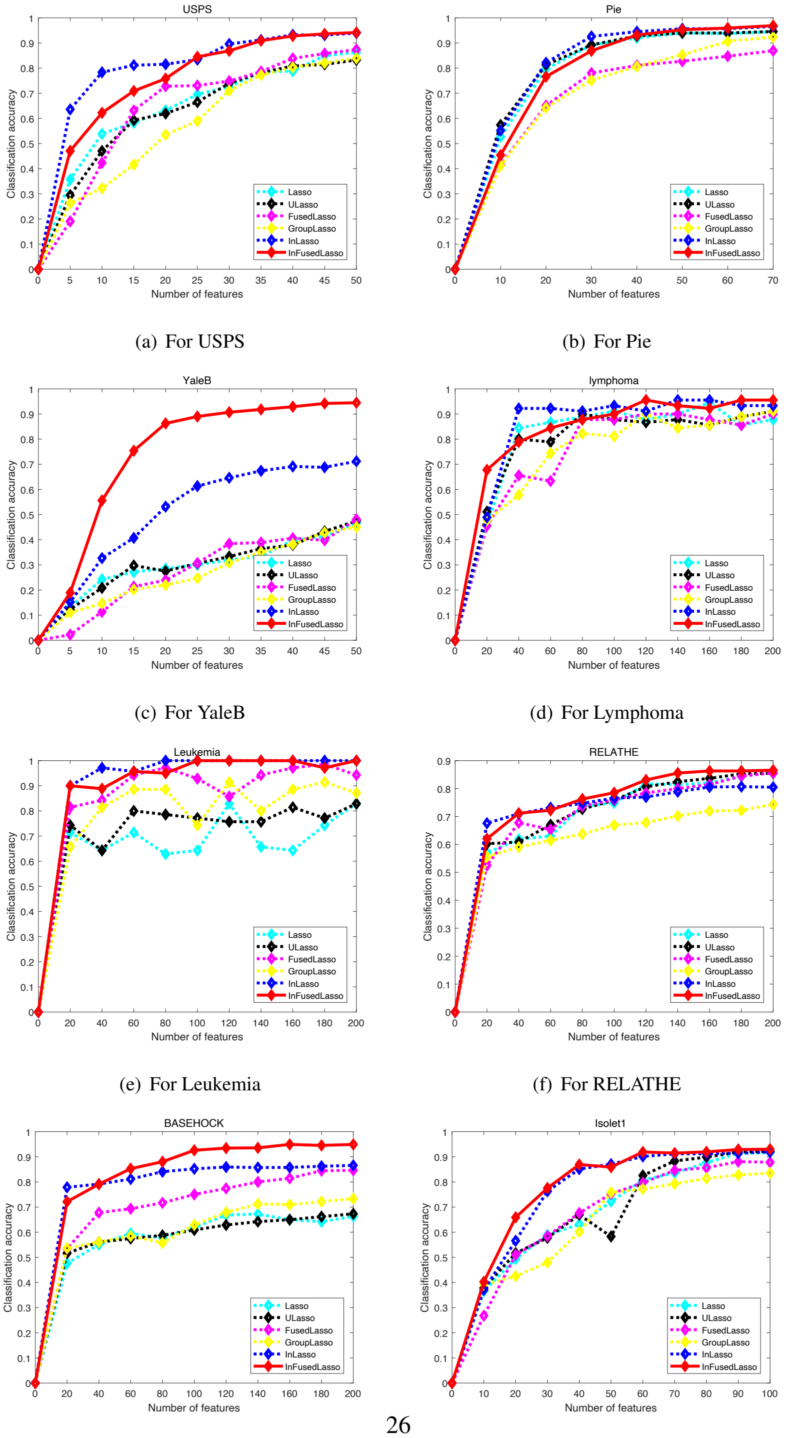
Table 6: InFusedLasso on representative datasets with high and low order interacted information.

Dataset	YaleB	Lymphoma	Leukemia	Isolet1
InFusedLasso (L)	94.94(50)	95.55(180)	98.57(40)	91.32(100)
InFusedLasso (K)	**96.37**(50)	**96.55**(120)	**100.00**(100)	**93.26**(100)

so, Fused Lasso, Group Lasso, InLasso, and InElastic Net. These experimental results indicate that the proposed InFusedLasso method can better learn the characteristics and interaction information residing on the features. This is because only the proposed approach can incorporate the structural information between feature samples through the structure-based feature graph representation.

5.3. Effects of the Kernel-based Modelling Method

To take our study one step further, we also compare the proposed framework associated with the distance-based graph features, denoted as InFusedLasso(D), that has been adopted by the previous Interacted ElasticNet method (InElaNet) [11]. Moreover, we also directly compare our proposed method to the Interacted ElasticNet method (InElaNet) [11], since this method can also encapsulate the structure correlated information. The results have been displayed in the Table 4. From this table, we can notice that the proposed method can outperform the InFusedLasso(D) method on all datasets. In addition, it is shown that the proposed InFusedLasso(K) method can outperform the InElaNet method on all of the datasets. The reason is that the required feature graph structures of the InFusedLasso(D) and InElaNet methods are computed based on the Eucliden distance. As we have stated earlier, the distance with large value may dominant the characteristics of the feature graph and influence the effectiveness. By contrast, the proposed InFusedLasso(K) method employs a new kernel-based graph modeling procedure to establish feature graphs and proposes a new structural interaction criterion to evaluate the joint relevancy of pairwise feature combinations in relation to the discrete target. As a result, the proposed InFusedLasso(K) method overcomes



(a) For USPS

(b) For Pie

(c) For YaleB

(d) For Lymphoma

(e) For Leukemia

(f) For RELATHE

(g) For BASEHOCK

(h) For Isolet1

Figure 4: Accuracy versus the number of selected features.

the shortcomings of the InFusedLasso(D) method and the InElaNet method. In addition, we observe that the InFusedLasso(D) method achieves better performance than InElaNet [11] on five out of the eight datasets, including YaleB, Lymphoma, Leukemia, RELATHE, and BASEHOCK. This is because the InFusedLasso(D) method is based on the Fused Lasso regularization terms, whereas the InElaNet method is based on the Elastic Net regularization terms. Although both methods are based on distance-based graph-features, the InFusedLasso(D) method can enhance the trade-off between relevancy of each individual feature and the redundancy of feature pairs, thus leading to better performance than the InElaNet method.

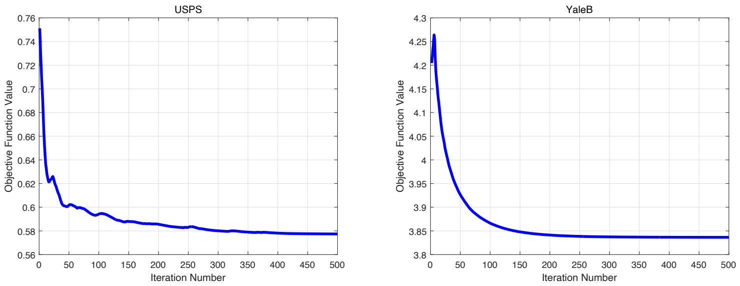
Overall, the experimental results verify that the proposed approach can locate more discriminative feature subsets than state-of-the-art feature selection approaches.

5.4. Exploiting the Ordering Relation

To evaluate the capability of fused lasso regularization term, i.e., the third regularization term in Eq.(8), on features with an ordering relation, we conduct experiments on four representative datasets. Because there is no natural ordering relation among the coefficients, we first calculate the individual relevance of each feature in relation to the target feature graph using the JSD measure for two probability distributions, and then rank these features according to their individual relevance scores. We compare the results obtained via the proposed method before and after this ranking procedure and display the results in Table 5. It is clearly shown in Table 5 that the proposed method achieves higher classification accuracies on all four datasets after ranking. This clearly demonstrates the advantage of Fused Lasso for features with ordering relation, that is, by reordering the features according to their relevance to the target feature, Fused Lasso can provide a better tradeoff between relevancy of each individual feature and the redundancy of pairwise features.

5.5. Comparison of high and low order interaction measures

To evaluate the effects of the proposed kernel-based interaction measure, which is a high-order interaction measure, with the low-order interaction measure proposed in [11], we compare the classification accuracy obtained via the proposed InFusedLasso



(a) For USPS

(b) For YaleB

Figure 5: Convergence curve for the optimization algorithm.

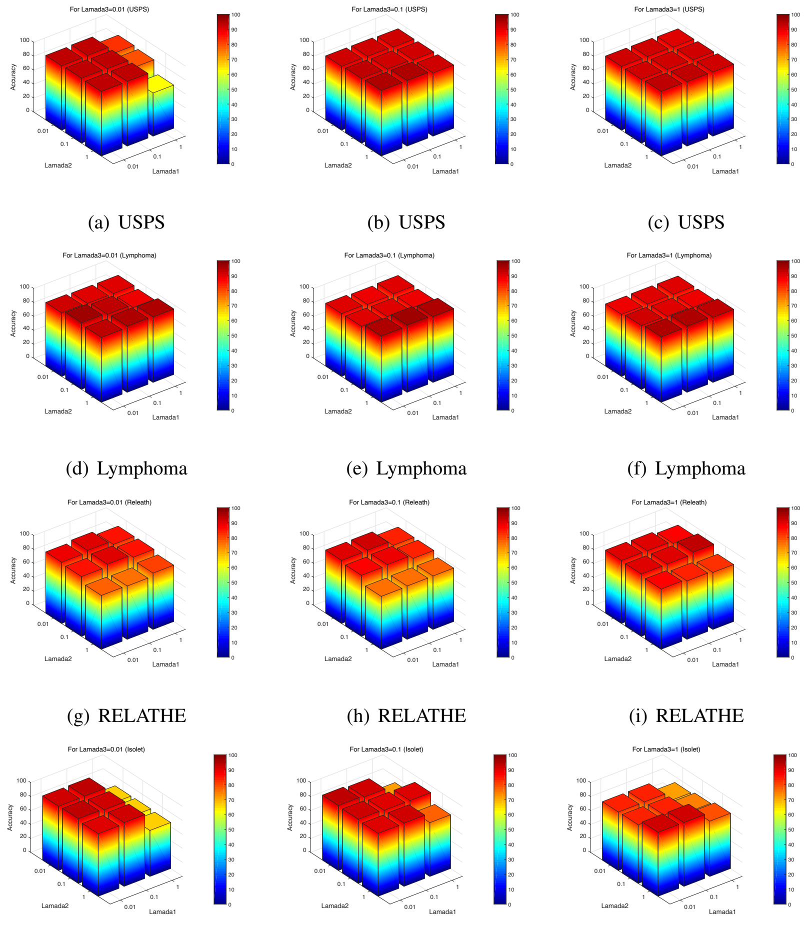
(K) method with InFusedLasso (L). Both methods are associated with the fused lasso regularization terms. The only difference is that the InFusedLasso (K) method utilizes the proposed high-order interaction measure and InFusedLasso (L) utilizes the interaction measure proposed in [11]. The classification accuracies are displayed in Table 6. It is clearly shown that the proposed kernel-based interaction measure has a positive impact on the performance of the proposed method.

5.6. Convergence Evaluation

In this subsection, we experimentally evaluate the convergence properties of the proposed optimization algorithm. Because we can observe similar results on all the datasets, we only display the convergence curves on two datasets, i.e., USPS and Yale-B. Specifically, the variations of the objective function values at each iteration are reported in Figure 5, which indicates that the proposed optimization algorithm converges as the iteration number within about 150 iterations, which ensures the efficiency and effectiveness of the proposed feature selection approach.

5.7. Parameter Sensitivity

Our proposed feature selection method consists of three adjustable parameters, i.e., λ_1, λ_2, and λ_3. Specifically, λ_1 and λ_2 are the tuning parameters for the Fused Lasso model, where λ_1 encourages the sparsity of β as in Lasso and λ_2 shrinks the differences between successive features specified in matrix $\mathbf{C}$ toward zero. Moreover, λ_3 is the corresponding tuning parameter of the structural interaction matrix $\mathbf{U}$. Different combinations of these three parameters might end with different classification results.



(a) USPS (b) USPS (c) USPS

(d) Lymphoma (e) Lymphoma (f) Lymphoma

(g) RELATHE (h) RELATHE (i) RELATHE

(j) Isolet1 (k) Isolet1 (l) Isolet1

Figure 6: Accuracies with different λ_1 and λ_2 when $\lambda_3 = 0.01, 0.1, 1.0$, respectively on USPS, Lymphoma, RELATHE and Isolet1 datasets.

In order to explore which combination of these three parameter values result in the best classification accuracy for a specific problem associated with the given classifier, we vary the value of λ_3 in the range of $0.01, 0.1, 1.0$ to investigate the benefits of the proposed structural interaction matrix. With a fixed value of λ_3, i.e., for $\lambda_3 = 0.01$, $\lambda_3 = 0.1$, $\lambda_3 = 1.0$, respectively, we vary the values of λ_1 and λ_2 in the range of $0.01, 0.1, 1.0$, and show the influence of the fused lasso term. The results are shown in Figure 6. More specifically, the first column of this figure corresponds to the results with different values for λ_1 and λ_2, with a fixed value of $\lambda_3 = 0.01$. The second column corresponds to the results with different values for λ_1 and λ_2, with a fixed value of $\lambda_3 = 0.1$. And the third column corresponds to the results with different values for λ_1 and λ_2, with a fixed value of $\lambda_3 = 1.0$. In addition, we choose the USPS, Lymphoma, RELATHE, and Isolet1 datasets for the parameter sensitivity analysis. The reasons for choosing these four datasets are as follows. First, USPS is a typical Image classification example with small number of features and large number of samples. Second, Lymphoma and RELATHE are typical examples from the Text classification and Biomedical classification tasks, and are both high in feature dimension and low in sample size. Third, Isolet1 is another example from the Speech classification task associated with small number of features and large number of samples.

From Figure 6, we have the following observations.

I. Generally, for each of the four datasets, 1) different combinations of the three parameters result in different classification accuracies; 2) there is a combination of the three parameters where the corresponding classification accuracy achieves the best; and 3) the values for the three parameters, in which the best classification accuracies are obtained, are different for the different datasets.

II. When the value of λ_3 is larger than both λ_1 and λ_2, the classification accuracies for the datasets tend to be higher. This indicates the proposed interaction matrix, which involves both feature redundancy and feature relevancy, has greater impact on the classification results than both the sparsity term and the differences between successive features. In addition, because the sparsity term and the differences between successive features represent information from the original feature space, this observation also reveals that the structural information encapsulated by the interaction matrix, has greater impact than the information from the original feature space.

III. With fixed values for λ_3 and λ_2, we can see that the larger the value for λ_1, the lower the classification accuracy is. This indicates that when choosing values for λ_1, we should keep its value lower than both λ_3 and λ_2.

6. Conclusions

In this paper, we have developed a new Fused Lasso model for feature selection. Unlike most state-of-the-art methods, our proposed approach incorporates structural information between pairwise samples into the feature selection process, which is significant for refining the performance of feature selection. More specifically, a new kernel-based similarity measure associated with the original Euclidean distance is proposed to construct the (target) feature graph structures. Furthermore, a new structural interaction measure associated with the feature graph representations is developed to simultaneously maximize joint relevancy of different pairwise feature combinations in relation to the target feature graphs and minimize redundancy among selected features. We embed the proposed interaction measure into a least square minimization model together with a Fused Lasso regularizer, which can enhance the trade-off between relevancy of each individual feature and the redundancy of pairwise features. Due to the nonseparability and nonsmoothness of the Fused Lasso regularization term in the objective function, an effective iterative algorithm is exploited to solve the proposed feature subset selection problem. Experiments demonstrate that the proposed feature selection approach is effective.

In future works, we may extend our approach associated with the quantum Jensen-Shannon divergence instead of the classical divergence measure. Specifically, in our previous works [38, 39], we have proposed a number of quantum Jensen-Shannon kernels using quantum walks. Since the quantum walks can encapsulate more complicated information of graph structures than the classical random walks used in this work. It would be interesting to extend the proposed feature selection method using the classical Jensen-Shannon divergence to that of using its quantum counterpart, resulting in a new quantum-based feature selection method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant no. 61976235 and 61602535), and the program for innovation research in Central University of Finance and Economics and the Youth Talent Development Support Program by Central University of Finance and Economics, No. QYP1908.

References

[1] S. Kashef, H. Nezamabadi-pour, A label-specific multi-label feature selection algorithm based on the pareto dominance concept, Pattern Recognit. 88 (2019) 654–667.

[2] W. Gao, L. Hu, P. Zhang, Class-specific mutual information variation for feature selection, Pattern Recognition 79 (2018) 328–339.

[3] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (6) (2018) 94:1–94:45.

[4] G. Ditzler, R. Polikar, G. Rosen, A sequential learning approach for scaling up filter-based feature subset selection, IEEE Trans. Neural Netw. Learning Syst. 29 (6) (2018) 2530–2544.

[5] M. A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000, 2000, pp. 359–366.

[6] G. Herman, B. Zhang, Y. Wang, G. Ye, F. Chen, Mutual information-based method for selecting informative feature sets, Pattern Recognition 46 (12) (2013) 3315–3327.

[7] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.

[8] H. Peng, F. Long, C. H. Q. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[9] M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, Expert Syst. Appl. 42 (22) (2015) 8520–8532.

[10] Z. Zhang, Y. Tian, L. Bai, J. Xiahou, E. R. Hancock, High-order covariate interacted lasso for feature selection, Pattern Recognition Letters 87 (2017) 139–146.

[11] L. Cui, L. Bai, Z. Zhang, Y. Wang, E. R. Hancock, Identifying the most informative features using a structurally interacting elastic net, Neurocomputing 336 (2019) 13–26.

[12] Y. Sun, H. J. Wang, M. Fuentes, Fused adaptive lasso for spatial and temporal quantile function estimation, Technometrics 58 (1) (2016) 127–137.

[13] R. Tibshirani, M. A. Saunders, S. Rosset, K. Knigh, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society Series B (Statistical Methodology) 67 (1) (2005) 91–108.

[14] J. W., L. J., L. N.A., L. J.M., Y. D., Some properties of generalized fused lasso and its applications to high dimensional data, Journal of the Korean Statistical Society 44 (3) (2015) 352–365.

[15] J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information, Neural Computing and Applications 24 (1) (2014) 175–186.

[16] F. Dornaika, Y. E. Traboulsi, Joint sparse graph and flexible embedding for graph-based semi-supervised learning, Neural Networks 114 (2019) 91–95.

[17] S. Liu, B. Wang, M. Xu, L. T. Yang, Evolving graph construction for successive recommendation in event-based social networks, Future Generation Comp. Syst. 96 (2019) 502–514.

[18] L. Bai, L. Cui, X. Bai, E. R. Hancock, Deep depth-based representations of graphs through deep learning networks, Neurocomputing 336 (2019) 3–12.

[19] Q. Gu, Z. Li, J. Han, Generalized fisher score for feature selection, in: UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011, 2011, pp. 266–273.

[20] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], 2005, pp. 507–514.

[21] M.Mandal, A.Mukhopadhyay, Unsupervised non-redundant feature selection: A graph-theoretic approach, in: FICTA2013,Proceedings of the International Conference on Frontiers of Intelligent Computing: Theorey and Applications, 2013, pp. 373–380.

[22] P. Moradi, M. Rostami, A graph theoretic approach for unsupervised feature selection, Eng. Appl. of AI 44 (2015) 33–45.

[23] Z. Zhang, E. R. Hancock, Hypergraph based information-theoretic feature selection, Pattern Recognition Letters 33 (15) (2012) 1991–1999.

[24] L. Cui, L. Bai, Y. Wang, X. Bai, Z. Zhang, E. R. Hancock, P2P lending analysis using the most relevant graph-based features, in: Proceedings of S+SSPR 2016, 2016, pp. 3–14.

[25] L. Cui, Y. Jiao, L. Bai, L. Rossi, E. R. Hancock, Adaptive feature selection based on the most informative graph-based features, in: Graph-Based Representations in Pattern Recognition - 11th IAPR-TC-15 International Workshop, GbRPR 2017, Anacapri, Italy, May 16-18, 2017, Proceedings, 2017, pp. 276–287.

[26] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58 (1) (1996) 267–288.

[27] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society 67 (5) (2005) 301–320.

[28] S. Ma, X. Song, J. Huang, Supervised group lasso with applications to microarray data analysis, BMC Bioinformatics 8.

[29] S. Chen, C. H. Q. Ding, B. Luo, Y. Xie, Uncorrelated lasso, in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA., 2013.

[30] B. Jiang, C. H. Q. Ding, B. Luo, Covariate-correlated lasso for feature selection, in: ECML PKDD 2014, Proceedings, Part I, 2014, pp. 595–606.

[31] J. Lin, Divergence measures based on the shannon entropy, IEEE Trans. Information Theory 37 (1) (1991) 145–151.

[32] L. Bai, L. Rossi, H. Bunke, E. R. Hancock, Attributed graph kernels using the jensen-tsallis q-differences, in: Proceedings of ECML-PKDD, 2014, pp. 99–114.

[33] G. Ye, X. Xie, Split bregman method for large scale fused lasso, Computational Statistics & Data Analysis 55 (4) (2011) 1552–1569.

[34] R. T. Rockafellar, A dual approach to solving nonlinear programming problems by unconstrained optimization, Math. Program. 5 (1) (1973) 354–373.

[35] T. R. Rockafellar, Convex Analysis, Princeton University Press, Princeton, NJ, 1997.

[36] J. Cai, S. J. Osher, Z. Shen, Split bregman methods and frame based image restoration, Multiscale Model. Simul. 8 (2) (2009) 337–369.

[37] Z. Zhang, L. Bai, Y. Liang, E. R. Hancock, Joint hypergraph learning and sparse regression for feature selection, Pattern Recognit. 63 (2017) 291–309.

[38] L. Bai, L. Rossi, A. Torsello, E. R. Hancock, A quantum jensen-shannon graph kernel for unattributed graphs, Pattern Recognit. 48 (2) (2015) 344–355.

[39] L. Bai, L. Rossi, L. Cui, Z. Zhang, P. Ren, X. Bai, E. R. Hancock, Quantum kernels for unattributed graphs using discrete-time quantum walks, Pattern Recognit. Lett. 87 (2017) 96–103.