Making Trillion Correlations Feasible in Feature Grouping and Selection

Yiteng Zhai, Yew-Soon Ong, *Senior Member, IEEE*, and Ivor W. Tsang

Abstract—Today, modern databases with "Big Dimensionality" are experiencing a growing trend. Existing approaches that require the calculations of pairwise feature correlations in their algorithmic designs have scored miserably on such databases, since computing the full correlation matrix (i.e., square of dimensionality in size) is computationally very intensive (i.e., million features would translate to trillion correlations). This poses a notable challenge that has received much lesser attention in the field of machine learning and data mining research. Thus, this paper presents a study to fill in this gap. Our findings on several established databases with big dimensionality across a wide spectrum of domains have indicated that an extremely small portion of the feature pairs contributes significantly to the underlying interactions and there exists feature groups that are highly correlated. Inspired by the intriguing observations, we introduce a novel learning approach that exploits the presence of sparse correlations for the efficient identifications of informative and correlated feature groups from big dimensional data that translates to a reduction in complexity from $O(m^2n)$ to $O(m \log m + \mathcal{K}_a mn)$, where $\mathcal{K}_a \ll \min(m, n)$ generally holds. In particular, our proposed approach considers an explicit incorporation of linear and nonlinear correlation measures as constraints in the learning model. An efficient embedded feature selection strategy, designed to filter out the large number of non-contributing correlations that could otherwise confuse the classifier while identifying the correlated and informative feature groups, forms one of the highlights of our approach. We also demonstrated the proposed method on one-class learning, where notable speedup can be observed when solving one-class problem on big dimensional data. Further, to identify robust informative features with minimal sampling bias, our feature selection strategy embeds the V-fold cross validation in the learning model, so as to seek for features that exhibit stable or consistent performance accuracy on multiple data folds. Extensive empirical studies on both synthetic and several real-world datasets comprising up to 30 million dimensions are subsequently conducted to assess and showcase the efficacy of the proposed approach.

Index Terms—Big dimensionality, feature grouping, sparse correlation, one-class learning, robust feature selection

1 Introduction

Feature correlation is among one of the most commonly used criteria of feature selection tasks in machine learning and data mining. While some researchers have focused on minimizing the correlations among features in the identified feature subset [1], [2], [3], others have exploited the mechanism of feature correlations via feature groups that capture new salient characteristics of the data [4], [5], [6]. These feature groups then serve as cues that could assist the human user in further analysis of the data. From a survey of the literature, feature correlation has been widely established as an important criterion for the identification of relevant, irrelevant, redundant and/or noisy features in learning and prediction tasks. It has received tremendous attentions over the past decades since datasets comprising large number of features are now becoming ubiquitous [2], [3], [4], [7].

Over the last decade, there has been an exponential growth in the dimensionality of the datasets that were generated. Such trends are non-isolated and can be observed across a plethora of diverse disciplines [8], [9], [10]. In bioinformatics, for instance, the search for a compact subset of relevant biomarkers from single-nucleotide polymorphism (SNP) in defining the behaviors of genes are now becoming prevalent [11]. Each SNP is often modelled as a feature and since each gene carries thousands of SNPs, the dimensionality of the genetic data easily reaches millions in size even though only a very small number of SNPs is relevant to the disease of interest [12]. Similar developments have been observed in the domain of imaging, where advancements in digital image sensors have given birth to digital cameras that can easily capture verisimilar photo with more than 41 megapixels. Notably, with a pixel-based feature representation, this translates to 41 million features in dimension per photo in deep learning framework. Likewise, in text mining, the feature space which is made up of unique words and/or phrases that appear in documents, tweet streams and webpages now easily extend to many millions of dimensions. Recently, the escalation of users that enjoy spending their leisure watching videos have propelled increasing research efforts towards intelligent data-centric media computing platforms. With the myriads of feature descriptors that are available for representing video contents (i.e., image, motion, acoustic and text, etc.), millions of features could easily transpire.

From our survey of the literature, today, modern databases with "Big Dimensionality" (i.e., millions of dimensions and above, as discussed in [8]) are becoming evident and such phenomenon will continue to be a growing trend.

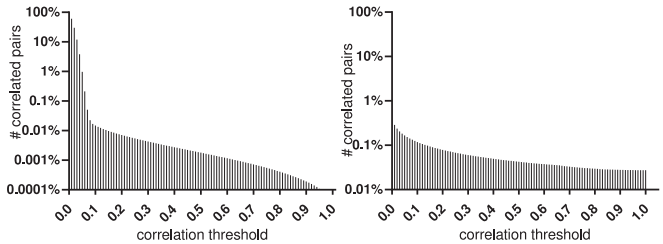
- *Y. Zhai and Y.-S. Ong are with the Rolls-Royce@NTU Corporate Lab, and the School of Computer Science and Engineering, Nanyang Technological University, Block N4, #2a-32, Nanyang Avenue, Singapore 639798. E-mail: {yzhai1, asysong}@ntu.edu.sg.*
- *I.W. Tsang is with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Sydney, N.S.W. 2007, Australia. E-mail: ivor.tsang@gmail.com.*

Manuscript received 7 Apr. 2014; revised 24 Nov. 2015; accepted 22 Jan. 2016. Date of publication 22 Feb. 2016; date of current version 10 Nov. 2016. Recommended for acceptance by K. Borgwardt.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2533384

For a detail exposition on the emerging phenomenon of big dimensionality, the reader is referred to [8]. As the dimensionality of datasets continues to push the capability limits of the algorithms, it is becoming clear that the complexity of the feature grouping and selection tasks being addressed began to overwhelm the algorithms available, i.e., due to the exponential increase in data dimension. In particular, existing approaches that require the calculations of pairwise correlations in their algorithmic designs cannot cope well with such high dimensional datasets elegantly and often scored miserably, since computing the full correlation matrix (i.e., square of dimensionality in size) can become computationally very intensive. Notably, a dataset with *millions* of features would translate to *trillions* of correlations to be computed. Although some works have been proposed on fast correlation findings [13], [14], [15], it is still worth noting that such degree of extreme computational complexity poses a challenge that has received much lesser attention in the field of machine learning and data mining research.

To reveal and illustrate the complexity of such a challenge, the efforts to compute the correlations of two commonly used and well established datasets are analyzed in what follows, including `psoriasis` with 529,651 SNPs and `news20.binary` with 1,355,191 word frequencies[1]. Theoretically, it can be asserted that for a simple brute force approach, a total number of $\binom{m}{2}$ computations would be necessary to obtain the pairwise correlations between all features in the datasets considered, wherein m denotes the number of features. In particular, **0.14 and 0.92 trillion** correlation computations[2] are necessary on these datasets. On the `psoriasis` dataset, which has only 529,651 features, it already took us 20.6 days of wall clock time to compute the full pairwise correlations of the feature sets in LIBSVM format, on an Intel Core i7-930 Processor. This clearly poses a serious impediment to the successful use of the feature correlation criterion on big dimensional datasets. Thus, there is a need for fresh computational and statistical learning paradigms to address such emerging challenge explicitly.

Fortunately, our detailed analyses on the well established datasets (which exhibit characteristics of big dimensionality) revealed that an extremely small portion (i.e., less than 0.1 percent for these two datasets considered) of the feature pairs have been found to be highly correlated. To illustrate this observation, we summarized the distributions of correlated feature pairs for the `psoriasis` and `news20.binary` datasets in Figs. 1(a) and 1(b), respectively. In the figure, each bar denotes the percentage of feature pairs (the y-axis) that satisfies a given correlation threshold interval (as indicated on the x-axis). From the figure, the percentage of feature pairs is noted to decrease exponentially for increasing correlation threshold values. To be precise, 99.985 percent of the feature pairs in `psoriasis` and 99.882 percent in the `news20.binary` dataset have correlation values lower than the threshold of 0.1. This implies that majority of the feature pairs are uncorrelated or the features are sparsely correlated. In this paper, we term this



(a) `psoriasis` (b) `news20.binary`

Fig. 1. Distributions of correlated feature pairs in some established datasets, wherein each bar denotes the percentage of feature pairs (the y-axis) that satisfies a given correlation threshold interval (the x-axis), i.e., $1 - \mathrm{CDF}_i$.

phenomenon as "sparse correlation" and our aspiration is to exploit this sparse correlation that is made available through the "Blessings of Big Dimensionality" [8].

In this paper, we introduce a novel learning approach that exploits the presence of sparse correlations for the efficient identifications of informative feature groups from datasets, especially in big dimensionality which involves general classification tasks including *binary classification* and *one-class learning problem*. Our proposed approach is a general feature grouping and selection framework, which considers an explicit incorporation of correlation measures as constraints in the learning model for different types of learning problem. An efficient embedded feature selection strategy, designed to remove large numbers of non-contributing correlations that could confuse the classifier, while identifying the informative feature groups, is then introduced. Extensive empirical studies on both synthetic and several real-world datasets comprising up to 30 million dimensions are subsequently conducted to assess and showcase the efficacy of our proposed approach. The core technical contributions of the current research work are summarized as follows:

1) The current work represents a first attempt to incorporate both linear and non-linear feature correlation measures for feature grouping and selection in binary and one-class machine learning settings. The inclusion of correlation constraints among features in the learning model facilitates possible identifications of informative feature groups.
2) To achieve the goal, we explicitly define the notions of *support feature* and *affiliated feature*. The former denotes the highly informative features with lower peer correlation, while the latter are features that are highly correlated to the support feature. Support-Affiliated Feature Groups are then established by an aggregation of the affiliated features that correspond to each support feature.
3) To identify the support-affiliated feature groups, in the proposed linear correlation setting, we derive a generalized relation between the absolute pairwise Pearson's correlation coefficient and the discriminative score of features. With such relationship established, the eliminations of uncorrelated feature pairs can thus be carried out without the need for a full correlation computation. It is worth noting that, this translates to a reduction in complexity on correlation computations, from $O(nm^2)$ to $O(m \log m + \mathcal{K}_a mn)$,

1. Note that, in our real-world experimental studies, besides these two, datasets with up to 30 million dimensions are considered.

2. To be exact, 140,264,826,075 and 918,270,645,645 correlation computations, respectively.

where m is the dimensionality, n is the size of data, $\mathcal{K}_a$ is the total number of support features identified and $\mathcal{K}_a \ll \min(m, n)$ generally holds.

4) To generalize the proposed framework over that proposed earlier in [4], further studies on i). generalizations of Proposition 1 by relaxing the assumption made on data normalization is proved; ii). nonlinear feature correlation (e.g., symmetrical uncertainty) is also considered; iii). the one-class learning setting, where notable acceleration over popular LIBSVM one-class SVM has been observed.
5) In addition, to reduce sampling bias caused by inadequate big dimensional training instances, we embed the V-fold cross validation as part of our feature selection scheme so as to converge to the set of robust features that exhibits stable prediction accuracy across multiple folds of data.

The remainder of this paper is organized as follows. In Section 2, some of the core definitions used in the paper are presented. Further, Section 3 introduces our proposed methodology to identify the support-affiliated feature groups effectively and efficiently. Section 4 gives a brief review of the related works on feature grouping. Then, we present the experimental setup and obtain results in Section 5. The conclusive remark and future work of interest are given in Section 6.

2 Preliminaries and Motivations

In this section, we present the core definitions and concepts that are used throughout the rest of the paper.

2.1 Definitions

In this paper, we define m as the dimensionality of the data, and n is the number of training data observations. $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ represents the intact training data, wherein each observation is denoted by $\mathbf{x}_i \in \mathbb{R}^m$, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$. Each vector $\mathbf{x}_i$ is associated with an output label $y_i \in \{\pm 1\}$ for binary-class problem, and $\mathbf{y}$ is defined as the vector of labels in the training data. Moreover, let $\mathbf{f}_j$ denote a row vector corresponding to the j^{th} feature of all observations in $\mathbf{X}$, thus $\mathbf{X} = [\mathbf{f}'_1, \ldots, \mathbf{f}'_m]' \in \mathbb{R}^{m \times n}$ holds. Additionally, the element-wise product between two matrices $\mathbf{A}$ and $\mathbf{B}$ is introduced as $\mathbf{A} \odot \mathbf{B}$, where $|\cdot|$ represents the cardinality operator unless specified. Symbols "$\mathbf{0}$" and "$\mathbf{1}$" are the column vectors comprising all zeros and all ones. And for each $\mathbf{f}$, the corresponding mean and standard deviation of the entries in $\mathbf{f}$ are indicated as $\mu_{\mathbf{f}}$ and $\sigma_{\mathbf{f}}$, respectively.

2.2 Feature Correlation Measures, *corr*$(\cdot, \cdot)$

In this section, we illustrate both the linear and nonlinear instantiations of $corr(\cdot, \cdot)$.

Amongst various correlation measures, Pearson's correlation coefficient (PCC) is one of the most commonly used *linear* correlation measure [16]. The PCC for a pair of features $\mathbf{f}_j$ and $\mathbf{f}_k$, $\rho(\mathbf{f}_j, \mathbf{f}_k)$, can be defined as follows:

$$corr_{\mathbf{linear}}(\mathbf{f}_j, \mathbf{f}_k) : \rho(\mathbf{f}_j, \mathbf{f}_k) = \frac{cov(\mathbf{f}_j, \mathbf{f}_k)}{\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k}}, = \frac{(\mathbf{f}_j - \mu_{\mathbf{f}_j}\mathbf{1}')(\mathbf{f}_k - \mu_{\mathbf{f}_k}\mathbf{1}')'}{n\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k}}, \quad (1)$$

wherein $cov(\mathbf{f}_j, \mathbf{f}_k)$ designates the covariance of the two features. However, as its polarity does not affect the informativeness of a selected feature, the coefficient is hereinafter referred to the absolute form in the present study.

From linear to *nonlinear* correlations, mutual information (MI) represents a well established measure for feature selection [17], which takes the form of $I(\mathbf{f}_j; \mathbf{f}_k) = H(\mathbf{f}_j) + H(\mathbf{f}_k) - H(\mathbf{f}_j, \mathbf{f}_k)$ (with $H(\cdot)$ denoting the entropy [18]) that measures the level of information sharing between feature $\mathbf{f}_j$ and $\mathbf{f}_k$ (i.e., $H(\mathbf{f}_j) \cap H(\mathbf{f}_k)$, where $H(\mathbf{f}) = -\sum_i p(f_i)\log_2 p(f_i)$). In feature selection, MI is typically used for assessing the ranking of the features in classification problem, i.e., a higher $I(\mathbf{f}; \mathbf{y})$ implies a higher devotion of feature $\mathbf{f}$ to class y.

MI is ranged as $[0, \infty]$ such that it is not a good quantization for pairwise correlation. Alternatively, the symmetrical uncertainty (SU) [19], which is a form of normalized MI has often been considered, which is defined by

$$corr_{\mathbf{nonlinear}}(\mathbf{f}_j, \mathbf{f}_k) : U(\mathbf{f}_j, \mathbf{f}_k) = \frac{2I(\mathbf{f}_j; \mathbf{f}_k)}{H(\mathbf{f}_j) + H(\mathbf{f}_k)}. \quad (2)$$

Note that, both absolute PCC ($|\rho(\cdot, \cdot)|$) and SU ($U(\cdot, \cdot)$) are symmetrical measures that lie in $[0, 1]$. Without loss of generality, other forms of correlation measure with a range of $[0, 1]$ may also apply in $corr(\cdot, \cdot)$. Further, a high (low) value of $corr(\cdot, \cdot)$ indicates that the pair of features considered are strongly (weakly) correlated. Hence if two features are fully independent, their correlation shall be 0. On the other hand, when they are completely correlated to each other, namely, one feature can exactly predict the other, 1 follows.

2.3 Support and Affiliated Features

The core objective of traditional feature selection approaches is to identify a reduced feature subset of informative features [3], [20]. In contrast to previous studies, this paper focuses on discovering the underlying interactions among informative features and capturing the salient characteristics within the data, based on the conception of sparse correlations and feature groupings, since such groupings can be useful to assist users in their interpretations of the data for further analysis. More specifically, we aim at identifying feature groups through pairwise feature correlation among informative features. Though several prior works [5], [6], [21] have highlighted the benefits of identifying feature groups (e.g., many biological studies have suggested that SNPs usually work in groups for some genetic activities), how to define the feature groups for general learning tasks is non-trivial.

In the present study, our interest is to identify a sparse feature subset of support features, while discovering the feature groups. Each feature group comprises a parent support feature with affiliated features as children that are strongly correlated to it. In what follows, we give the definitions of the support feature and affiliated feature, which form the basis of the current work.

Definition 1. *A* Support Feature *(SF_k) denotes the most informative (discriminative) feature w.r.t. the output labels among the residual features. All of the support features identified, as depicted in full circles of Fig. 2, are uncorrelated or weakly correlated to one another.*

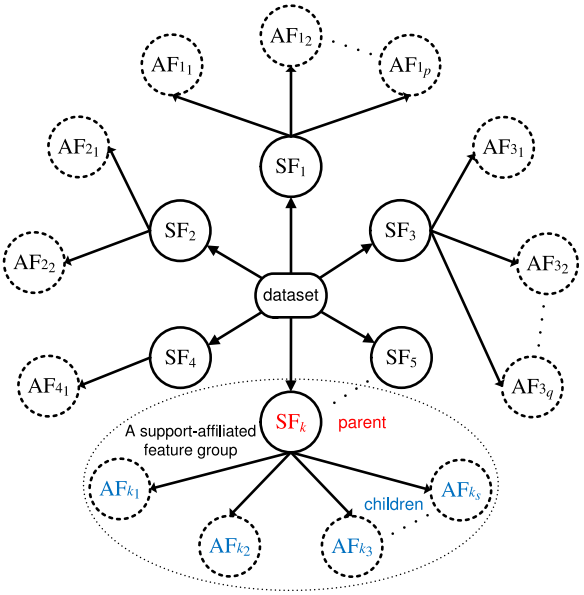


Fig. 2. Structural relationship of support-affiliated feature groups (denoted using dotted ellipse). SF_k: support feature (parent denoted using full circle), AF_{k_s}: affiliated features (children of SF_k as denoted by dotted circles).

Definition 2. *An* Affiliated Feature *(AF_{k_s}) should also be an informative feature, which shares similar predictive capability with the associated support feature SF_k, and is strongly correlated with SF_k (i.e., $corr(SF_k, AF_{k_s}) \geq \varepsilon$). The dotted circles of Fig. 2 showcase this type of features.*

With the above definitions, the interest of our current work is to discover support-affiliated feature groups that takes the form of Fig. 2 from various datasets in different machine learning settings.

3 Group Discovery Machine

In this section, a novel feature grouping and selection method, labeled here as the *group discovery machine* (GDM), is introduced for the discovery of support-affiliated feature groups in various machine learning tasks, such as one-class and two-class problems. The essential backbone of the GDM is a sparse SVM with an efficient quadratically constrained quadratic programming (QCQP) solver. We introduce an explicit incorporation of the pairwise linear/nonlinear correlation measures as constraints in the learning model to discover the appropriate support-affiliated feature groups.

3.1 General Correlation Constraints

Similar to the idea of feature indicator in [22], a vector $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_m]' \in \{0,1\}^m$ is introduced to define whether a corresponding SF is selected ($\delta_j = 1$) or not ($\delta_j = 0$), such that the decision function is given by: $f(\mathbf{x}) = \mathbf{w}'(\mathbf{x} \odot \boldsymbol{\delta})$, where the vector $\mathbf{w} \in \mathbb{R}^m$ denotes weight vector. To limit the number of selected features to be lower than $\mathcal{K}_a$, the ℓ_0-constraint $\|\boldsymbol{\delta}\|_0 \leq \mathcal{K}_a$ is imposed for the purpose of feature selection. Further, to constrain the correlation among the selected features, the following constraint on $\boldsymbol{\delta}$ is explicitly introduced here as

$$\delta_j \delta_k = 0 \text{ if } |corr(\mathbf{f}_j, \mathbf{f}_k)| \geq 1 - \tau, \forall j, k \text{ with } j \neq k. \quad (3)$$

With this constraint, the feature pair is regarded as uncorrelated if their correlation coefficient falls below the bound $(1-\tau)$, where $\tau \in [0, 0.5]$. Next we define $\boldsymbol{\Delta} = \{\boldsymbol{\delta} | \sum_{j=1}^m \delta_j \leq \mathcal{K}_a;\ \delta_j \in \{0,1\};\ \delta_j \delta_k = 0 \text{ if } |corr(\mathbf{f}_j, \mathbf{f}_k)| \geq 1-\tau, \forall j, k \text{ with } j \neq k\}$ as the domain for $\boldsymbol{\delta}$. Further, (3) explicitly defines $\binom{m}{2} \Rightarrow O(m^2)$ quadratic constraints with m numbers of integer variables. As previously noted, our present task is inclined to solve problems with *millions of dimensions*, which translates to *trillion quadratic constraints*. Moreover, seeking the solution $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ involves a process of combinatorial subset selection, resulting in extremely high computational cost, especially on big dimensional data. In what follows, we describe in detail our proposed approach to deal with the trillion correlation constraints that arise.

3.2 Proposed Formulation

In GDM, the interest is to find a large margin decision function $f(\mathbf{x})$ for robust prediction, and seamlessly identify the informative yet uncorrelated feature subset that satisfies the constraints defined in (3). For the purpose of simplicity, the square hinge loss in SVM is considered, thus arriving at the following optimization problems:

$$\begin{aligned} \min_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \min_{\mathbf{w}, \gamma, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 - \gamma + \frac{C}{2}\sum_{i=1}^n \xi_i^2 \\ s.t. \quad & y_i \mathbf{w}'(\mathbf{x}_i \odot \boldsymbol{\delta}) \geq \gamma - \xi_i \quad i = 1, \ldots, n, \end{aligned} \quad (4)$$

where $\xi_i \geq 0$ is the slack variable, $\gamma/\|\mathbf{w}\|$ denotes the margin and C is a tradeoff parameter to regulate the function complexity $\|\mathbf{w}\|_2^2$ and the training error (ξ_i's). Note, as discussed earlier, the optimization problem in (4) with constraints defined in (3) is a challenging problem, as a result of the explosion in the number of constraints involving big dimensional data.

3.3 Solving the Problem Iteratively with Cutting Plane Algorithm

Cutting planes are a major component of the mixed integer linear optimization solver for accelerating the progress by removing fractional solutions. Recently, the *cutting plane algorithm* has reported much success in many problems involving vast varieties of constraints, including SVM training [23], structure prediction [23], maximum margin clustering [24] and so on.

Taking the cue, here we solve problem (4) by incorporating a cutting plane approach. To begin, the inner minimization section of problem (4) considers a dual form of SVM w.r.t. $\mathbf{w}, \gamma$ and ξ_i. Thus (4) becomes a minimax saddle-point problem. Inspired by applying the minimax optimization theory, a tight convex relaxation to problem (4) can be attained, which takes the form of a QCQP problem:

$$\begin{aligned} & \min_{\boldsymbol{\alpha} \in \mathcal{A}, \theta} \theta : \theta \geq g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}), \forall \boldsymbol{\delta} \in \boldsymbol{\Delta} \quad \text{or} \quad \min_{\boldsymbol{\alpha} \in \mathcal{A}} \max_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \\ & \text{define} \quad g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) = \frac{1}{2}\left\| \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta}) \right\|^2 + \frac{1}{2C}\boldsymbol{\alpha}'\boldsymbol{\alpha}, \end{aligned} \quad (5)$$

wherein $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]'$ is the vector of dual variables, $\mathcal{A} = \{\boldsymbol{\alpha} | \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0, \forall i = 1, \ldots, n\}$ defines the domain of $\boldsymbol{\alpha}$, and θ is the upper bound of $g_{\boldsymbol{\delta}}(\cdot)$. Nevertheless, since

there are as many as $(\sum_{i=0}^{\mathcal{K}_a}\binom{m}{i})$ quadratic constraints in problem (5), the problem remains to be plagued with computational complexity issues. The *cutting-set methods* developed in [25] considers a general worst-case convex optimization problem with arbitrary dependence on the uncertain parameters. Hence, rather than solving the original problem which involves a vast number of constraints, cutting-set is used here to generate a subset of active constraints in an iterative manner. This leads to a relaxed optimization problem with the current constraint set considered. At the t^{th} iteration, $\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \geq g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}), \forall \boldsymbol{\delta}^t \in \boldsymbol{\Delta}$ holds, and correspondingly $\boldsymbol{\delta}^t$ is constrained by a $\mathcal{K}_b$ (i.e., support feature size per iteration), where $\sum_t \mathcal{K}_b^t = \mathcal{K}_a$ generally holds. Thus, for a reduced active constraint set $\Lambda \subset \boldsymbol{\Delta}$, the lower bound approximation of (5) can be obtained as $\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} g_{\boldsymbol{\delta}}(\boldsymbol{\alpha}) \geq \max_{t=1,\ldots,T} g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha})$ with $T = |\Lambda|$, where T is the maximum number of constraints (iterations) imposed. This leads to solving a reduced problem of (5) that takes the form

$$\min_{\boldsymbol{\alpha}\in\mathcal{A},\theta} \theta \ : \ \theta \geq g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}), \quad \forall\ \boldsymbol{\delta}^t \in \Lambda. \tag{6}$$

3.4 Training with Multiple Kernel Learning

In this section, MKL optimization technique is considered for solving the problem defined in (6), wherein the aim is to jointly learn both the kernel and SVM parameters, or briefly, to identify the most appropriate kernel for addressing the task on hand [26].

Since problem (6) follows a convex QCQP problem, we introduce μ^t as the dual variable of each constraint. The Lagrangian function then takes the form of

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}) = -\theta + \sum_{t, \boldsymbol{\delta}^t\in\Lambda} \mu^t(\theta - g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha})). \tag{7}$$

Setting the derivative w.r.t. θ as zero, $\sum \mu^t = 1$ can be attained. We set $\boldsymbol{\mu}$ as the vector of μ^t's, and $\mathcal{U} = \{\boldsymbol{\mu} | \sum \mu^t = 1, \mu^t \geq 0\}$ defines the domain of $\boldsymbol{\mu}$. Consequently, the Lagrangian function $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu})$ can be rewritten as

$$\begin{aligned} &\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\mu}\in\mathcal{U}} \sum_{\boldsymbol{\delta}^t\in\Lambda} -\mu^t g_{\boldsymbol{\delta}^t}(\boldsymbol{\alpha}) \\ &= \min_{\boldsymbol{\mu}\in\mathcal{U}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}(\boldsymbol{\alpha}\odot\mathbf{y})'\left(\sum_{\boldsymbol{\delta}^t\in\Lambda} \mu^t \mathbf{X}_t\mathbf{X}_t' + \frac{1}{C}\mathbf{I}\right)(\boldsymbol{\alpha}\odot\mathbf{y}), \end{aligned} \tag{8}$$

where $\mathbf{X}_t = [\mathbf{x}_1 \odot \boldsymbol{\delta}^t, \ldots, \mathbf{x}_n \odot \boldsymbol{\delta}^t]'$, and the equation follows on account of the fact that the objective function is concave in $\boldsymbol{\alpha}$ and convex in $\boldsymbol{\mu}$. The recently developed MKL is ideal for solving the resultant minimax problem (8) [26], [27], where the kernel matrix $\sum_{\boldsymbol{\delta}^t\in\Lambda} \mu^t \mathbf{X}_t\mathbf{X}_t'$ to be learnt is a convex combination comprising $|\Lambda|$ number of base kernel matrices $(\mathbf{X}_t\mathbf{X}_t')$, each of which is constructed from a feasible $\boldsymbol{\delta}^t \in \Lambda$.

To summarize, the steps for solving the proposed problem are outlined in Algorithm 1, wherein some of the notations are explained thereinafter. Specifically, for each iteration of Algorithm 1, one needs to figure out the worst case analysis (i.e., finding the most violated constraint $\boldsymbol{\delta}^t$) of Problem (5) [25], [28], which is described in the following Sections 3.5, 3.6, and 3.7. The obtained $\boldsymbol{\delta}^t$ is then appended into the active constraint set Λ, which forms a subset of $\boldsymbol{\Delta}$. Last but not least, the problem w.r.t. Λ can be solved via efficient QCQP solvers [4].

Algorithm 1. Group Discovery Machine—GDM($\mathbf{w}, \boldsymbol{\delta}, \mathcal{D}$)

Input: Dataset $\mathcal{D}(\mathbf{X}, \mathbf{y})$, zero-one vector $\boldsymbol{\delta} \in \mathbb{R}^m$, support feature size per iteration $\mathcal{K}_b$ and correlation threshold τ.
Output: Index set $\mathcal{SF}$ for SFs and $\mathcal{AF}$ for AFs.
Initialization: $\boldsymbol{\alpha} = \mathbf{1}/n, \boldsymbol{\delta} = \mathbf{0}^m, \mathcal{S} = \emptyset$ and $\mathcal{Q} = \emptyset$.
for $t = 1$ **to** T **do**
1: Call $\boldsymbol{\delta}^t = \text{CRM}(\mathcal{D}, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}, \mathcal{AF})$.
2: Set $\Lambda = \Lambda \cup \{\boldsymbol{\delta}^t\}$ and solve (6), while updating $\boldsymbol{\alpha}^{t+1}$.
3: Quit if the objective value is convergent.
end for

3.5 Correlation Redundancy Matching (CRM): Finding the Most Violated Constraints $\boldsymbol{\delta}$

In this section, the worst case analysis of problem (5), which plays a key role in *Cutting Plane Algorithm* [25] is presented. In the current problem setting, problem (6) is transformed into solving the following integer optimization problems:

$$\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \left\| \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta}) \right\|^2. \tag{9}$$

In general, solving such a problem is considered NP-hard. However, since one can obtain $\| \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \odot \boldsymbol{\delta})\|^2 = \| \sum_{i=1}^n (\alpha_i y_i \mathbf{x}_i) \odot \boldsymbol{\delta}\|^2 = \sum_{j=1}^m s_j^2 \delta_j$, where we define s_j as the *feature discriminative score* that follows

$$s_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij} = \sum_{i=1}^{n} \alpha_i y_i f_{ji} = \mathbf{f}_j \tilde{\boldsymbol{\alpha}}, \tag{10}$$

with $\tilde{\boldsymbol{\alpha}} = [\alpha_1 y_1, \ldots, \alpha_n y_n]'$, indicating that the informative features should accord with features of largest absolute value feature score $|s_j|$'s. Moreover, recall that we embed correlation measures in $\boldsymbol{\delta}$, thus a natural question arises: considering all the correlated features, which one poses higher importance to the output labels?

To address this question, first of all, it is necessary to offer the instantiations of SF and AF in the proposed GDM. As discussed previously, SFs refer to the most informative features with relatively low pairwise correlations in this work. AFs, on the other hand, refer to the correlated features associated with each SF correspondingly. The parent-child structured relationship between SFs and AFs is illustrated in Fig. 2.

Definition 3. SF and AF in GDM. *Given any exemplar vector $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^n$ and a collection of feature vectors $\{\mathbf{f}_i\}$, where $\mathbf{f}_i' \in \mathbb{R}^n$. The SF is given by $\max_i |\mathbf{f}_i \tilde{\boldsymbol{\alpha}}|$ for the given $\tilde{\boldsymbol{\alpha}}$. The remaining correlated features in $\{\mathbf{f}_j\}$ w.r.t. $\mathbf{f}_i$ (with $|corr(\mathbf{f}_i, \mathbf{f}_j)| \geq 1 - \tau$) then denote the AFs.*

For the sake of conciseness, let $\mathcal{SF}$ be the index set of the SFs and here we introduce a data structure $\mathcal{AF} = \{\mathcal{G}_j\}$ to represent the hierarchical structure of features, where $\mathcal{G}_j$ denotes the index set of the AFs for the j^{th} SF. In this manner, all the correlated features can be identified and

archived instead of omitting them.[3] Moreover, based on the definitions above, once a support feature (SF_j) is identified (i.e., the feature with the largest $|s_j|$), all relevant features (AF_{j_s}) that correlate with SF_j then become the affiliated features that correspond to it. As the present proposed method discovers the correlated feature groups, it is labelled as the GDM here. Note that, alternatively, one could employ a brute-force approach to search across all features and pairwise correlations to identify all feature groups that achieves the similar goal. However, such a scheme [2] can be computationally intensive even with small dataset and would become computational intractable on big dimensional data. For the details on the intensiveness of a brute-force approach, the reader is referred to the Appendix E, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2016.2533384.

Seeking for the most violated constraints $\boldsymbol{\delta}$ is then termed here as Correlation Redundancy Matching (CRM) procedure. In CRM, once an SF is identified, the AFs are isolated from the rest of the features based on their correlations w.r.t. the SF. The above procedure is repeated until a maximum of $\mathcal{K}_b$ unique support features are identified in each iteration.

3.6 CRM with Linear Correlation $corr_{\text{linear}}(\cdot,\cdot)$

In this section, we illustrate the way of feature grouping with linear correlation, i.e., using PCC $\rho(\cdot,\cdot)$ as the correlation measure. To this end, we begin with a proposition to prove the case of linear correlation, which serves as a generalization of that previously presented in [4] on assumption made pertaining to data normalization: for a group of strongly correlated features, if one of them is informative to the output labels, all of them can be treated identically (i.e., all of them will make positive contributions to the output label).

Proposition 1. *Given a nonzero column vector $\tilde{\boldsymbol{\alpha}}$ and any two feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$, suppose their absolute PCC $|\rho(\mathbf{f}_1,\mathbf{f}_2)| \geq (1-\tau)$, then $|\,|\mathbf{f}_1\tilde{\boldsymbol{\alpha}}| - |\mathbf{f}_2\tilde{\boldsymbol{\alpha}}|\,| \leq \sqrt{n(\Delta_\sigma^2+\Delta_\mu^2+2\tau\sigma_{\mathbf{f}_1}\sigma_{\mathbf{f}_2})}\|\tilde{\boldsymbol{\alpha}}\|$ holds, where the $\tau \in [0,1]$ and $\Delta_\sigma = \sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2}$ while $\Delta_\mu = \mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2}$.*

Proof. The proof is particularized in Appendix A, available in the online supplemental material. □

The above results state that if two feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ are highly correlated, their distance measure (or correlation) to any exemplar vector $\tilde{\boldsymbol{\alpha}}'$ will be close to one another. In what follows, we present a theorem to illustrate that in practice one can address the linear pairwise feature correlation by scanning only a small subset of the features on big dimensional problems.

Theorem 1. *Given a nonzero column vector $\tilde{\boldsymbol{\alpha}}$ and any two feature vectors $\mathbf{f}_j$ and $\mathbf{f}_k$, suppose $|\rho(\mathbf{f}_j,\mathbf{f}_k)| \geq (1-\tau)$ and $\mathbf{f}_j$ is the support feature (i.e., $\mathbf{f}_k$ is qualified as an affiliated feature of $\mathbf{f}_j$) with feature score $|s_j| = |\mathbf{f}_j\tilde{\boldsymbol{\alpha}}|$ based on Equation (10), then the feature score of $\mathbf{f}_k$ satisfies $|s_k| \geq |s_j| - \sqrt{n(\Delta_\sigma^2+\Delta_\mu^2+2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$ with $\Delta_\sigma = \sigma_{\mathbf{f}_1} - \sigma_{\mathbf{f}_2}$ and $\Delta_\mu = \mu_{\mathbf{f}_1} - \mu_{\mathbf{f}_2}$.*

3. The practice of existing works in the literature is to omit all correlated features, i.e., redundancy reduction.

Proof. From Proposition 1, we can obtain that $|\,|\mathbf{f}_j\tilde{\boldsymbol{\alpha}}| - |\mathbf{f}_k\tilde{\boldsymbol{\alpha}}|\,| \leq \sqrt{n(\Delta_\sigma^2+\Delta_\mu^2+2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$, under $|\rho(\mathbf{f}_j,\mathbf{f}_k)| \geq 1-\tau$. Further, since $\mathbf{f}_j$ is the support feature, which has the highest feature score among all other features, so $|\mathbf{f}_j\tilde{\boldsymbol{\alpha}}| \geq |\mathbf{f}_k\tilde{\boldsymbol{\alpha}}|$. Correspondingly, we have $|\mathbf{f}_k\tilde{\boldsymbol{\alpha}}| = |s_k| \geq |s_j| - \sqrt{n(\Delta_\sigma^2+\Delta_\mu^2+2\tau\sigma_{\mathbf{f}_j}\sigma_{\mathbf{f}_k})}\|\tilde{\boldsymbol{\alpha}}\|$. This completes the proof. □

The above theorem states that if two features are strongly correlated, their scores will be close to one another. In other words, for a given support feature $\mathbf{f}_j$, all other features with scores that fall below the arrived bound at the correlation level of $(1-\tau)$, shall not be considered as the affiliated features of $\mathbf{f}_j$. Correspondingly, this facilitates possible eliminations of vast numbers of uncorrelated features without the need to undergo extensive correlation computations. The details of seeking $\boldsymbol{\delta}$ using PCC is then illustrated in Algorithm 2, and we term GDM that employs the PCC as GDM-PCC.

Algorithm 2. CRM$(\mathcal{D},\mathcal{K}_b,\tau,\boldsymbol{\alpha}^t,\mathcal{SF},\mathcal{AF})$ with PCC

1: Initialize $k=1$ and denote $\varrho\|\tilde{\boldsymbol{\alpha}}\|$ as the bound arrived from Theorem 1. Set the output $\boldsymbol{\delta}^t = \mathbf{0}$.
2: Compute feature score vector $\mathbf{s}$ according to (10) and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$.
while $\|\boldsymbol{\delta}^t\|_0 < \mathcal{K}_b$ **do**
  Pick the k^{th} feature $\mathbf{f}_z$ from $\mathcal{D}$, where $z = \mathcal{E}(k)$
  if $(|\,|s_z|^t - |s_j|^t\,| > \varrho_{z,j}\|\tilde{\boldsymbol{\alpha}}\|^t$ with all existed SF $\mathbf{f}_j$) **then**
    $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
  else
    For the SFs that satisfy $(|\,|s_z|^t - |s_j|^t\,| \leq \varrho_{z,j}\|\tilde{\boldsymbol{\alpha}}\|^t)$, compute $\rho(\mathbf{f}_z,\mathbf{f}_j)$.
    if $(\exists j, \rho(\mathbf{f}_z,\mathbf{f}_j) \geq 1-\tau)$ **then**
      $\mathcal{AF}.\mathcal{G}_j = \mathcal{AF}.\mathcal{G}_j \cup \{z\}$ ($\mathbf{f}_z$ is set as new AF for SF_j)
    else
      $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
    end if
  end if
  Set $k = k+1$
end while

3.7 CRM with Nonlinear Correlation $corr_{\text{nonlinear}}(\cdot,\cdot)$

Besides linear correlation, nonlinear relationship is also considered as fundamental to many statistical, physical and biological phenomena [29]. Among the nonlinear correlation measures, the normalized mutual information is considered as an important criterion for pairwise vectors [30]. Here we show that, GDM offers flexibility and room for nonlinear correlation measure to handle more complex tasks. For the sake of brevity, in this section, we consider the SU as the normalized MI in GDM (i.e., $corr_{\text{nonlinear}}(\cdot) = U(\cdot)$) and term it GDM-SU correspondingly. Algorithm 3 summarizes the pseudo code of feature grouping and selection with nonlinear correlation SU.

3.8 GDM on One-Class Learning (GDM-OC)

In this section, we further illustrate the generality of the proposed GDM framework for solving *One-Class Learning* problems. Particularly, many applications, including fraud detection and novelty detection, are commonly seen as one-class learning problems. Under such a problem setting, an expected classification is taken to

differentiate between known objects (i.e., the target class) from unknown objects (i.e., outlier, abnormal observation), while preserving the corresponding distribution of the known. Therefore, data with single class label can be used to assess whether a learning machine is able to properly preserve the boundary of the learnt class. In what follows, we show that the proposed GDM framework can readily accommodate one-class problem with ease by holding the label information of each observation (i.e., $y_i = 1$ for all known objects) in the proposed sparse SVM formulation,

Algorithm 3. CRM$(\mathcal{D}, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}, \mathcal{AF})$ with SU

1: Initialize $k = 1$ and set the output $\boldsymbol{\delta}^t = \mathbf{0}$.
2: Compute feature score vector $\mathbf{s}$ and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$.
while $\|\boldsymbol{\delta}^t\|_0 < \mathcal{K}_b$ **do**
  Pick the k^{th} feature $\mathbf{f}_z$ from $\mathcal{D}$, where $z = \mathcal{E}(k)$
  if $(\exists j, U(\mathbf{f}_z, \mathbf{f}_j) \geq 1 - \tau)$ for SF_j **then**
    $\mathcal{AF}.\mathcal{G}_j = \mathcal{AF}.\mathcal{G}_j \cup \{z\}$ ($\mathbf{f}_z$ is set as new AF for SF_j)
  else
    $\mathcal{SF} = \mathcal{SF} \cup \{z\}$ and $\delta_z^t = 1$ ($\mathbf{f}_z$ is set as new SF)
  end if
  Set $k = k + 1$
end while

$$\min_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \min_{\mathbf{w},\gamma,\xi} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 - \gamma + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \qquad s.t. \quad \mathbf{w}'(\mathbf{x}_i \odot \boldsymbol{\delta}) \geq \gamma - \xi_i \quad i = 1, \ldots, n. \tag{11}$$

Taking the dual form, similar derivation to Sections 3.3, 3.4, and 3.5 can be attained with the exception of $g_{\boldsymbol{\delta}}(\boldsymbol{\alpha})$ in problem (5) becoming $\frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i(\mathbf{x}_i \odot \boldsymbol{\delta})\right\|^2 + \frac{1}{2C}\boldsymbol{\alpha}'\boldsymbol{\alpha}$, while problem (8) takes the form of

$$\min_{\boldsymbol{\mu}\in\mathcal{U}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{\boldsymbol{\delta}^t\in\Lambda}\mu^t\mathbf{X}_t\mathbf{X}_t' + \frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha}. \tag{12}$$

The corresponding feature discriminative score can then be computed with $s_j = \sum_{i=1}^{n}\alpha_i x_{ij} = \sum_{i=1}^{n}\alpha_i f_{ji} = \mathbf{f}_j\boldsymbol{\alpha}$. At the same time, the decision to predict normal patterns can be determined by $f(\mathbf{x}) = \text{sgn}((\mathbf{w}' \odot \boldsymbol{\delta})\mathbf{x} - \gamma)$, where γ is the learnt threshold. In this case, Algorithm 1 remains to hold (unless all inputs share the same class), and γ can be obtained from $\boldsymbol{\alpha}$ from the dual solution based on Karush-Kuhn-Tucker (KKT) conditions [31].

Besides this, one can also adapt the constraint to perform different learning tasks, such as classification on structured or semi-structured data [32] or using the selected features for graph embedding [33].

3.9 Convergence Analysis of GDM

In this section, we conduct the convergence analysis by introducing some theorems and propositions.

Theorem 2. *Let $(\boldsymbol{\alpha}^*, \theta^*)$ be the global optimal pair of (5), define*

$$\beta^k = \max_{1\leq i\leq k} g_{\boldsymbol{\delta}^i}(\boldsymbol{\alpha}^k) = \min_{\boldsymbol{\alpha}\in\mathcal{A}} \max_{1\leq i\leq k} g_{\boldsymbol{\delta}^i}(\boldsymbol{\alpha}) \quad \text{and} \quad \varphi^k = \min_{1\leq j\leq k} g_{\boldsymbol{\delta}^{j+1}}(\boldsymbol{\alpha}^j),$$

where k denotes the number of iterations, then we have $\beta^k \leq \theta^* \leq \varphi^k$. And with an increasing k, β^k is monotonically increasing while φ^k is monotonically decreasing.

Proof. The proof is particularized in Appendix B, available in the online supplemental material. □

Proposition 2. *With the proposed* correlation redundancy matching *algorithm, the most violated constraint selection problem (9) can be solved with the most informative feature selected.*

Proof. The proof is particularized in Appendix C, available in the online supplemental material. □

The next theorem indicates that the proposed GDM can globally converge and exhibits the non-monotonic property for feature selection.

Theorem 3. *For each iteration of Algorithm 1, suppose that the reduced minimax subproblem (6) can be globally solved and the most violated constraint selection (9) can be solved with the most informative feature selected, Algorithm 1 terminates after a finite number of iterations.*

Proof. The proof is particularized in Appendix D, available in the online supplemental material. □

In general, the *Cutting Plane Algorithm* typically converges to robust optimal solution within tens of iterations under the worst case analysis (i.e., finding the most violated constraint $\boldsymbol{\delta}^t$), where notable performances on many real applications have been reported [25].

3.10 Implementation Issue

It is often the case that big dimensional datasets are usually sparse. To store such massive data, a sparse format is preferable. For example, LIBSVM has implemented a data structure with double linked list structure, where indices are generated for fast accessing of each sample point. However, the LIBSVM format[4] requires a sequential search to retrieval the features, it is inefficient for feature selection involving big dimensional datasets. To facilitate fast direct access of features, in this work, we have introduced a feature-based indexing for each feature. With such an implementation, the efficiency is significantly improved as demonstrated in Appendix E, available in the online supplemental material.

3.11 Complexity Analysis

In GDM, the most violated $\boldsymbol{\delta}$ is obtained via the CRM algorithm, where the m number of features are firstly sorted based on the feature score metric $|s_j|$ in GDM. Consequently, $\mathcal{K}_b$ number of most informative features with correlation consideration (i.e., w.r.t. the other predictive features) are identified. For T iterations, there will be $T \times \mathcal{K}_b$ SFs at most, which is the worst case to consider in MKL. Nevertheless, as previously discussed, the cutting plane strategy requires a small T for convergence to happen—a cap of ten iterations is used in the experimental study on binary classification, thus T is not a crucial term in the complexity. For the sake of brevity, the detailed complexity analysis on the

4. The comparison of GDM and baseline with both implementations are presented in Appendix E, available in the online supplemental material.

TABLE 1
Complexity Analysis for Each Iteration in GDM

Sub-procedure		Complexity
CRM	Feature Score **s**	$O(mS)$
	Sort $\lvert s_j\rvert$'s	$O(m \log m)$
	Feature Grouping	$O(\mathcal{K}_b mn)$
MKL		$O(T\mathcal{K}_b n)$

two iterative steps of the proposed method is given in Table 1, where S indicates the number of support vectors in the SVM. Note that for the feature grouping stage of CRM, the complexity $O(\mathcal{K}_b mn)$ is for the worst case, however, due to the phenomena of "sparse correlation" highlighted in the Introduction, the true complexity is much less than $O(\mathcal{K}_b mn)$. To conclude, GDM is very efficient for the real-world data with "sparse correlation".

4 Related Works on Feature Grouping

In the recent years, feature grouping, has been shown to be a good means for building simple structure of the data. A feature group accumulates substantive characteristics of the features (i.e., it has a functional interpretation to the prediction task). Furthermore, considering each group as a branch, a feature structure of the data can be established by an aggregation of the branches. The established feature structure can be helpful for grasping the properties of the feature space of interest and assist users in their interpretations of the data for further analysis. The graph-guided fused lasso (GFlasso) is among the early feature grouping approaches and operates by identifying feature groups based on the graph-structure defined over the features [5]. Moreover, GFlasso employs a sparse regularization over a graph to penalize the differences in feature coefficients β_i and β_j by $|\beta_i - \text{sign}(\rho_{i,j})\beta_j|$, and then connects/associates $\mathbf{f}_i$ to $\mathbf{f}_j$ when $\rho_{i,j} > 0$. Other works include the Elastic-Net [34] and group Lasso [35], wherein the former uses the hybridization of ℓ_1 and ℓ_2 regularizers to gather strongly correlated features into groups when dealing with high dimensional problems; the latter, however, introduces an extension of the lasso penalty, which is deemed as an intermediate between ℓ_1 and ℓ_2 penalty, so as to favor robust features. Octagonal shrinkage and clustering algorithm for regression (OSCAR), on the other hand, incorporates the ℓ_∞-penalty as a means to reduce similar feature pairs [36], while the ℓ_1 regularizer is maintained for feature selection purposes. Further, Zhong and Kwok introduced an efficient projection step to accelerate the process of feature grouping [37]. Recently, Yang et al. employed a convex function to penalize the pairwise infinity norm of connected regression/classification coefficients, while achieving simultaneous feature grouping and selection [6]. They considered a non-convex optimization function to enforce bias alleviation.

In spite of the increasing efforts that focus on identifying feature groups, existing strategies have met with limited success on Big Volume and big dimensional data [8]. The key factor responsible for this lies in the need to compute an extremely large covariance/correlation matrix for big dimensional data, which is computationally intractable. Notably, a dataset with millions of features translates to trillions of correlations to be computed. In contrast to previous works, with the proposed GDM, we exploited the presence of sparse correlations for the efficient identifications of informative and correlated feature groups from big dimensional datasets, without the need to compute a full covariance/correlation matrix.

5 Experimental Study

In this section, we present the experimental study on the proposed GDM-PCC and GDM-SU together with several state-of-the-art feature selection methods, including: 1) mRMR[5] [2], 2) FCBF[6] [3], 3) RCFS [38], 4) SVM-RFE [20] and 5) ℓ_1-SVM[7] [39]. In addition, some state-of-the-art feature grouping methods are also considered here to assess the performance efficacies of GDM. These include: 6) OSCAR [37], 7) ncFGS & ncTFGS [6] and 8) GFlasso [5]. To provide insights to the contributions of incorporating correlation constraints, $\tau = 0$ is configured to arrive at GDM$_{\tau=0}$. This forms the baseline where no differentiations between support and affiliated features are made (i.e., features with highest $|s_j|$'s are preferred).

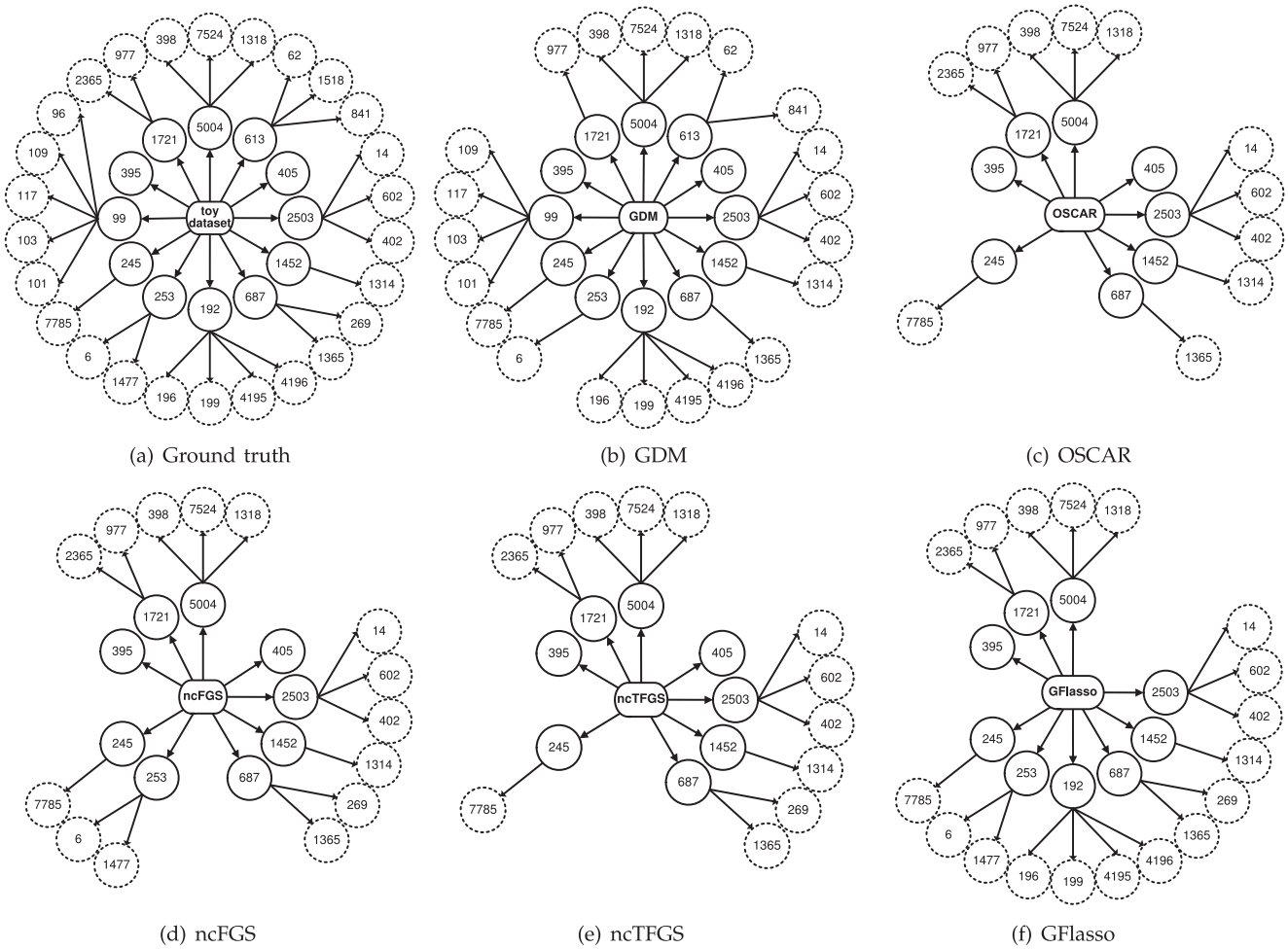
5.1 Experimental Setup

Among the feature selection methods considered here, mRMR, FCBF and RCFS represent filter methods, SVM-RFE is a representative of the wrapper method, while ℓ_1-SVM belongs to the family of embedded method. For the feature grouping methods, OSCAR, ncFGS & ncTFGS and GFlasso all operate based on the strategy of pruning the covariance/correlation matrix. The configurations of all the methods considered are set to be consistent to those used in the respective articles reported, and implemented in C++. Moreover, to facilitate fair comparisons, the standard SVM classifier is used as the underlying classifier. In GDM, the parameter C of the standard SVM is set to '1', while in ℓ_1-SVM, C is unique and vary with the number of features selected. In the experimental study, we set the correlation parameter as $\tau = 0.3$ for GDM-PCC while $\tau = 0.4$ for GDM-SU, and for the mutual information calculation in GDM-SU, we use 10 percent of the feature value range as the estimator. To show the results of different numbers of selected features, $\mathcal{K}_b$ is set as natural number (e.g., 1, 2, ..., 10). All experiments are conducted on the PC with Intel Xeon CPU E5-2695 v2 (2.4 GHz, 2 processors) and 128GB memory under Windows Server 2012 R2 Standard.

5.2 Results on Synthetic Dataset

To illustrate the mechanisms of the proposed GDM, we begin our study on a synthetic dataset, where the ground truth support-affiliated relationships are known in advance. Correspondingly, the aim is to verify if our proposed method is able to discover the feature groups and how it fares against the existing state-of-the-art feature grouping methods. The training set comprises 2,048 observations, each having 10,000 features. There are 12 predefined informative features which have been further expanded as

5. http://penglab.janelia.org/proj/mRMR
6. http://www.cs.man.ac.uk/~gbrown/fstoolbox
7. http://www.csie.ntu.edu.tw/~cjlin/liblinear



(a) Ground truth (b) GDM (c) OSCAR

(d) ncFGS (e) ncTFGS (f) GFlasso

Fig. 3. Feature group structures generated by the various feature grouping methods.

12 feature groups with variant size, as depicted in Fig. 3(a). Each support feature has 0 to 5 affiliated features as children, while the others then form the noisy features. The predictive ability of each group is configured to follow a normal distribution $\mathcal{N}(0,1)$. The pseudo algorithm of generating this dataset is provided in Appendix G, available in the online supplemental material, and since we use linear correlation to generate the feature group, the GDM-PCC is adopted in the experiment. Thus, in what follows, GDM indicates GDM-PCC in this section.

The experimental results obtained by the various feature grouping methods and feature selection methods on the synthetic dataset are tabulated in Table 2. In particular, success hits and training time are the performance metrics used to assess the algorithms under consideration as summarized in the table. *Success Hits* provides a measure on the completeness of a feature grouping method or feature selection method in correctly identifying all the core features. *Training Time*, on the other hand, gives the wall-clock time incurred to train a learning model. From the results in Table 2, both the wrapper (i.e., SVM-RFE) and embedded methods (i.e., GDM, ℓ_1-SVM) have been observed to attain competitive performances on both metrics for feature selection methods.

With the correlation constraint in (3) disabled by setting $\tau = 0$, $\text{GDM}_{\tau=0}$ is observed to achieve performances that are close to the ℓ_1-SVM. This is unsurprising due to the similar sparse SVM strategy used in both algorithms. Filter methods such as mRMR and FCBF suffered the worst

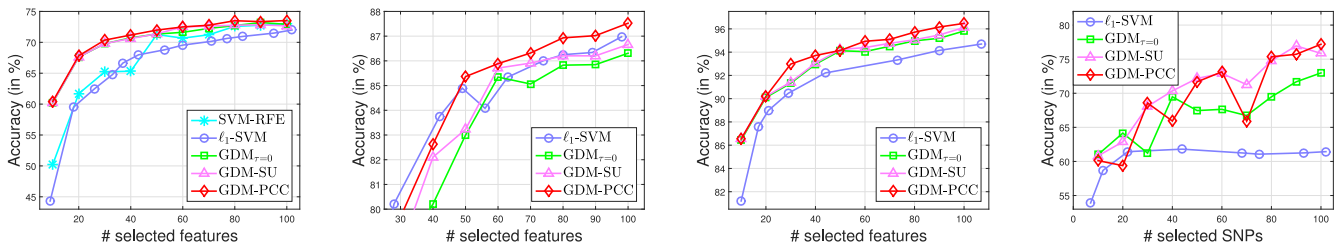
TABLE 2
Results on Synthetic Dataset of Various Methods

Methods	Feature Grouping					Feature Selection					
	GDM	OSCAR	ncFGS	ncTFGS	GFlasso	$\text{GDM}_{\tau=0}$	mRMR	FCBF	RCFS	SVM-RFE	ℓ_1-SVM
Success Hits	**86.8 percent**	50.0 percent	60.5 percent	52.6 percent	71.1 percent	**9/3/0**	2/3/7	6/1/5	7/2/3	8/3/1	**10/2/0**
Training Time	0.85 ± 0.13	130.68	455.34	456.62	132.49	$\mathbf{0.31 \pm 0.16}$	2.28 ± 0.25	4.56 ± 0.52	101.95	33.73 ± 0.98	$\mathbf{0.33} \pm 0.07$

Success Hits *stands for the completeness in identifying the features. It measures the matching degree for feature grouping methods (i.e.,* $\frac{\#\text{CORRECT}}{\#\text{ALL}} \times 100\%$*) whilst taking the form of "# correct SF/# correct AF/# incorrect feature" for feature selection methods. The* Training Time *is in reported seconds, wherein the deviation below 1 second is reported.*

TABLE 3
Characteristics of the Real-World Datasets Considered

Dataset	# Features	# Training	# Positive	# Negative	# Testing	# Nonzeros	Density	Size on disk
news20.binary	1,355,191	9,996	6,000	3,996	10,000	3,584,383	2.646e-04	133.52 MB
kdd2010	**29,890,095**	**19,264,097**	16,579,660	2,684,437	748,401	585,609,985	1.017e-06	4.96 GB
webspam	16,609,143	280,000	169,786	110,214	70,000	1,044,482,369	2.245e-04	**23.31** GB
psoriasis	529,651	2,176	1,131	1,045	545	**1,424,895,775**	**0.989**	11.67 GB



(a) news20.binary (b) kdd2010 (c) webspam (d) psoriasis

Fig. 4. Testing accuracy (in percent) on real-world datasets.

performances. Nevertheless, RCFS, though achieved competitive result, incurred a learning time of 102 seconds, which is 300+ times more than embedded feature selection method, i.e., GDM$_{\tau=0}$ and ℓ_1-SVM.

The feature group structures generated by the various feature grouping methods are then depicted in Figs. 3(b), 3(c), 3(d), 3(e), and 3(f). Visually, Fig. 3(b) which is the group structure produced by the proposed GDM, is noted to match the ground-truth feature groups of Fig. 3(a) most closely than all the others, (i.e., Figs. 3(c), 3(d), 3(e), and 3(f)). To assert this quantitatively, we use *Success Hits* to measure the obtained feature group structures of the various feature grouping methods (i.e., matching degree for feature grouping methods, $\frac{\#\text{CORRECT}}{\#\text{ALL}} \times 100\%$)), among which GDM ranked at the top, with a value of 86.84 percent (i.e., 33/38 of ground truth features). GFlasso managed to uncover 27/38 of the ground truth features, while all the other feature grouping methods only identified less than 23 of the ground truth features. For the sake of illustration, the important features that have been identified by the feature grouping methods are depicted in Fig. 3.

5.3 Results on Real-World Datasets

To assess the performance of GDM on real world settings, here we present a study on a range of big dimensional datasets from diverse domains. The first is the 20 newsgroup dataset, which has been size-balanced for binary text classification with each class comprising 10 groups and labelled here as news20.binary[8]. The second is the kdd2010[8] challenge dataset used in the educational data mining competition. The aim of the competition is to provide predictions on the "correct first attempt" for a subset of "steps". The third is the spam web page data webspam[9], which is collected by "Webb Spam Corpus 2006" with adequate number of spam pages for the purpose of spam detection. One biology data considered here is the psoriasis dataset comprising SNPs as the features. This data is collected from a collaborative association study of psoriasis (CASP)[10] to identify susceptibility pathways and important genes [40]. As SNPs data is very dense, this makes feature grouping and selection a challenging and non-trivial task.

To proceed with our study on the webspam and psoriasis datasets, we randomly select 80 percent of the entire observations as the training set, and the rest 20 percent are kept for testing. For news20.binary, we set the training and testing set to be equal in size due to the sparseness of the dataset, so as to better preserve the original feature space. Further, detailed information on the datasets is listed in Table 3, wherein the bold font indicates the core challenge of each dataset considered, e.g., the challenge of big dimensionality in kdd2010 with nearly 30 million features, high density characteristic of psoriasis and enormous storage requirement of webspam, etc.

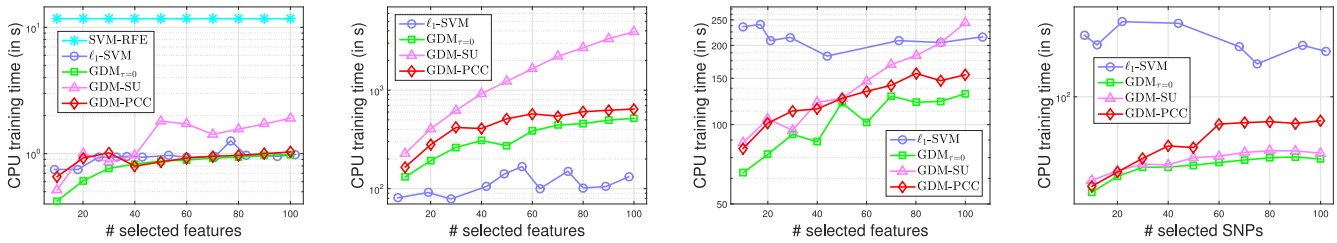
To adapt with the feature selection task, we use SFs to represent GDM series methods, i.e., GDM-PCC and GDM-SU. Further, to evaluate the feature selection performances, *classification accuracy*, *training time complexity* and *redundancy rate*[11] [41] are considered.

5.3.1 Accuracy Results

Fig. 4 summarizes the accuracy performance attained by the methods considered including SVM-RFE, ℓ_1-SVM, GDM$_{\tau=0}$, GDMs (GDMs refers to both GDM-SU and GDM-PCC in this section). Note that, all the other feature grouping methods as well as the filter feature selection methods have been observed to be inadequate for handling such high dimensionality considered, hence only results of the wrapper and embedded methods are reported. Furthermore, SVM-RFE consumed a large number of inner SVM evaluations, thus it fails to converge well on the three larger datasets under the limited training budget available.

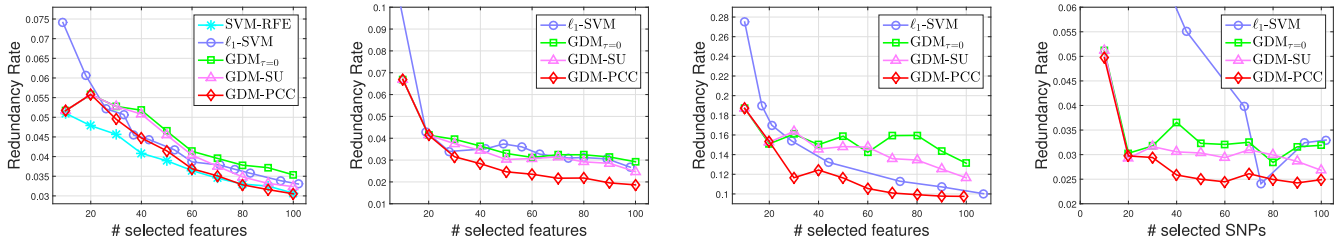
Overall, GDM attains very high accuracy improvements over the other methods on the real world datasets. From

8. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets
9. http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html
10. http://www.sph.umich.edu/csg/abecasis/casp
11. Redundancy rate generally assesses the averaged correlation among all the selected feature pairs.



(a) news20.binary (b) kdd2010 (c) webspam (d) psoriasis

Fig. 5. Training time (in seconds) on real-world datasets (in logarithmic scale, averaged from 5 runs).



(a) news20.binary (b) kdd2010 (c) webspam (d) psoriasis

Fig. 6. Redundancy rate on real-world datasets.

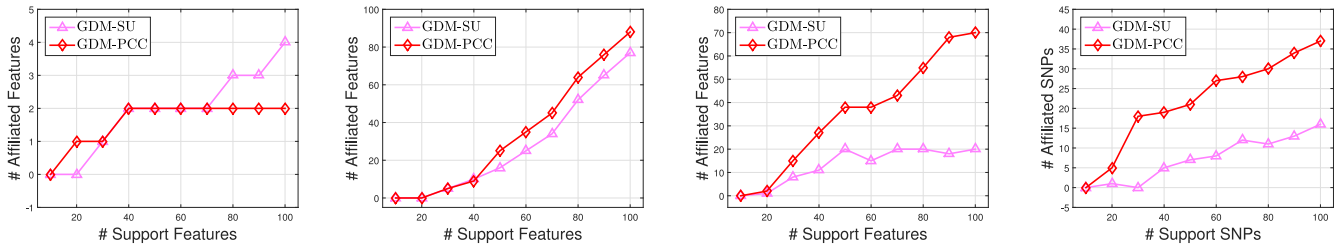
Fig. 4, we can also observe that both GDM-PCC and GDM-SU fare significantly better than GDM$_{\tau=0}$ on the larger datasets. This improvement can be attributed to the benefits brought about by the feature correlation constraints considered in GDM, i.e., identifying the SFs and AFs, since the only disparity between GDMs and GDM$_{\tau=0}$ lies in the lack of differentiations between SFs and AFs in the latter. While for news20.binary, since the informative features are mostly uncorrelated, only a small number of feature groups are identified by both GDM methods, hence similar accuracies are reported by GDMs and GDM$_{\tau=0}$. Besides, GDM-PCC performs slightly better than GDM-SU on three of the datasets, while on the psoriasis biological data, the nonlinear correlation measure of GDM-SU exhibited improved and more stable accuracy, especially when the number of selected SFs is small. Moreover, both GDMs and GDM$_{\tau=0}$ showcase statistically significant improvements in accuracy over the ℓ_1-SVM, however, relatively high variance is displayed in the prediction accuracies reported w.r.t. the increasing number of selected SNPs, which is mainly due to the *high dimension small sample size* characteristic of the psoriasis dataset that we will illustrate further in a section of further study (see Section 5.4).

5.3.2 Training Time Results

Fig. 5 further summarizes the training cost incurred, wherein embedded methods are noted to be more efficient on the news20.binary dataset when compared to the wrapper method, SVM-RFE. Moreover, Fig. 5(b) shows that, on the highly sparse kdd2010 dataset (see the density in Table 3), ℓ_1-SVM was able to take advantage on the sparseness characteristics of the kdd2010 dataset to achieve the shortest training time observed. However, on dense datasets, i.e., psoriasis, ℓ_1-SVM did not fare well and in fact incurred large training costs due to the inadequate management of memory resources. When no differentiations between SF and AF are considered, GDM$_{\tau=0}$ displays remarkable training efficiency compared to GDM-PCC and GDM-SU, as shown in Figs. 5(b) and 5(c), with some tradeoff in prediction accuracies. Most importantly, GDM$_{\tau=0}$ could possibly satisfy the real-time requirement of some applications if some form of parallel computing is used. Despite having to handle the massive number of feature correlation computations involving big dimensional dataset, GDM-PCC reported a comparable or smaller training time costs than ℓ_1-SVM. Further, referring to the number of selected SFs and AFs reported in Fig. 7, it can be observed that when the quantity of AFs in GDM-PCC and GDM-SU are similar, and GDM-SU incurs a higher computational time than GDM-PCC (e.g., as observed on the news20.binary and kdd2010 dasasets). However, on the dense dataset psoriasis, GDM-SU exhibits higher efficiency with a smaller identified feature groups (note that for webspam, GDM-PCC may take advantage of the sparseness in the dataset to reduce the cost of PCC calculation).

5.3.3 Redundancy Rate Results

Last but not least, in Fig. 6, the redundancy rate attained by various methods are depicted, wherein GDM-PCC achieves a low rate in most cases. Compared to embedded methods, both GDM-SU and GDM-PCC outperform the ℓ_1-SVM in terms of redundancy reduction, while exhibiting improved accuracy performance at the same time. Moreover, this implies that the SFs selected by GDM-SU or GDM-PCC approximately form a *good feature set* [1]. Specifically, from Fig. 6(a), SVM-RFE also reported low redundancy rate. However, tracing back to Fig. 4(a), SVM-RFE did not fare so well on the accuracy performance metric. It appears that the redundancy of SVM-RFE suffers from the presence of irrelevant features. Overall, considering both the results in Figs. 4 and 6, both GDM-PCC and GDM-SU are noted to attain low redundancy rate and superior accuracy performance simultaneously on the real-world datasets and relative to all the feature selection methods considered.



(a) news20.binary (b) kdd2010 (c) webspam (d) psoriasis

Fig. 7. Number of AFs selected w.r.t. the number of SFs by GDM-PCC and GDM-SU.

5.4 Further Study: Embedded Cross Validation

As a result observed from Fig. 4, the accuracies of GDM-PCC grow smoothly on three of the real-world datasets, however, shake sharply on the psoriasis dataset, which is possibly due to the *high dimension small sample size* characteristic of the data. In practice, when dealing with big dimensional data, the feature size sometimes far exceeds the number of data samples by many orders of magnitude [42], [43]. In such cases, it is typical that any slight variations in the training data often produces radical changes in the selected feature models. In order to identify an intrinsic set of features that represent the entire data or the original feature space, here we aggregate the feature selection results of multiple feature selectors by means of voting/averaging, in the spirit of the ensemble feature selection strategy [42]. With a consensus made using diverse feature selectors, possible biases caused by the inconsistent distributions of the data can be reduced.

In achieving such a goal, the universal support feature set $\boldsymbol{\delta}^{uni}$ is elected, which is formulated as

$$\min_{\mathbf{w}^v, \boldsymbol{\delta}^{uni}} \sum_v \text{GDM}(\mathbf{w}^v, \boldsymbol{\delta}^{uni}, \mathcal{D}^v), \tag{13}$$

wherein the superscript $\cdot^v$ in the function denotes the v^{th} feature selector, hence $\mathbf{w}^v$ represents the corresponding weight vector and $\mathcal{D}^v$ the data subset. Further, a universal feature set $\boldsymbol{\delta}^{uni}$ is enforced to identify only single set of robust SFs kept in testing (i.e., rather than one set per cross validation run). To solve problem (13), we embed the traditional V-fold cross-validation within the GDM. Consequently, this framework is labelled here as Embedded Cross Validation (ECV), whose algorithm is summarized in Algorithm 4. Further, with such adoption, the complexity only involves the CRM V times, while exhibiting only one robust feature set.

We thus train this data using ECV, where $V = 5$ folds are adopt in the procedure. We can observe from Fig. 8 that ECV not only outperforms GDM in terms of prediction accuracy, it also reports a significantly more stable results. This underlines the benefits of embedding the cross validation with GDM to arrive at ECV that helps alleviate the bias of such data distribution, so as to converge to robust features. Thus, it is worth emphasizing that the aggregation of performances by diverse feature selector is helpful for improving the robustness of the selected features, especially in *high dimension small sample size* problem settings [8], [44], wherein different feature subsets are often reported to yield similar performances.

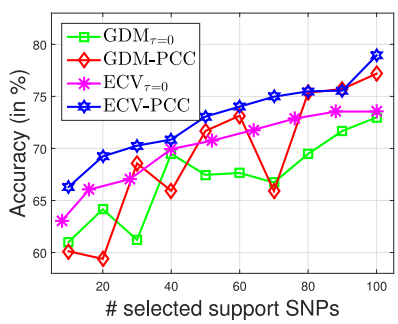


Fig. 8. Accuracy results on psoriasis for ECV.

Algorithm 4. GDM with Embedded Cross-Validation

Input: Dataset $\mathcal{D}(\mathbf{X}, \mathbf{y})$, $\mathcal{K}_b$, τ and number of fold V.
Output: Index sets of $\mathcal{SF}$ and $\mathcal{AF}$.
Initialize $\boldsymbol{\alpha} = \mathbf{1}/n, \mathcal{SF} = \emptyset$ and $\mathcal{AF} = \emptyset$, and randomly split the training set into V equal partitions $\mathcal{D}_v$.
for $t = 1$ **to** T **do**
  Compute feature score vector $\mathbf{s}$ and sort $|s_j|$ in descending order, record the feature ranking list as $\mathcal{E}$
  for $v = 1$ **to** V **do**
    Call $\boldsymbol{\delta}_v^t = \text{CRM}(\mathcal{D}_v, \mathcal{K}_b, \tau, \boldsymbol{\alpha}^t, \mathcal{SF}_v, \mathcal{Q}_v)$
  end for
  1: Set '1' for the indices of top $\mathcal{K}_b$ features from $\sum_v \boldsymbol{\delta}_v^t$ in $\boldsymbol{\delta}^t$.
  2: Archive $\mathcal{AF}$ from each $\mathcal{AF}_v$ that corresponding to $\mathcal{SF}$.
  3: Set $\Lambda = \Lambda \cup \{\boldsymbol{\delta}^t\}$ while updating $\boldsymbol{\alpha}^{t+1}$.
end for

5.5 Result on One-Class Study of GDM

In this section, we showcase the effectiveness of the proposed GDM-OC for solving one-class learning problem. The standard one-class SVM (OCSVM) integrated in

TABLE 4
Comparison of One-Class Learning Result between LIBSVM-OCSVM and GDM

Dataset	Accuracy		Training Cost (in sec)	
	OCSVM	GDM-OC	OCSVM	GDM-OC
news20.binary	**63.11** percent $\nu = 0.5$	62.90 percent 55 features	54.71	**1.37**
kdd2010	N. A. N. A.	**71.76** percent 20 features	N. A.	**408.49**
webspam	74.14 percent $\nu = 0.5$	**76.00** percent 70 features	1.75×10^5	**170.91**
psoriasis	**52.84** percent $\nu = 0.8$	48.26 percent 100 features	1,260.25	**89.49**

Fig. 9. Interpretability of selected support-affiliated feature groups.

LIBSVM[12] [45] is then considered for comparison. In the experimental study, all four real-world datatsets are considered to maintain consistency. The difference is that only the positive training points of the datasets are used as training data in the one-class setting. Different from binary-class setting, since the convergence preference is needed, 0.01 percent in precision of objective difference is set (In this case, $T = \infty$ in Algorithm 1). $\mathcal{K}_b$ is then set as natural number {2, 5, 10} and a better result is recorded. Further, the ν in LIBSVM-OCSVM is set from 0.2 to 0.8 with an interval of 0.1 and the best result for this method is presented for comparison. The empirical results obtained including accuracy and training cost are given in Table 4, from where one can easily figure out that with the embedded feature selection, not only is the training accuracy maintained, the training process is also accelerated. Notably, this speed up is in proportional to the density of the data, as can be observed from Table 3. This indicates that GDM suits sparse dataset very well.

6 Future Work and Conclusion

In this section, we begin with some future work discussion and follow up with the conclusion for the paper.

6.1 Future Work on the Benefits of Affiliated Features on Image Data

Image clustering is one of the most challenging tasks of computer vision research, especially when dealing with face images. The reason falls on the fact that very often the description of different faces can be rather similar whilst being very distinct only in the background. Thus, traditional clustering methods without proper feature normalization tend to be easily misled. Recently, Nie et al. presented a spectral embedded clustering (SEC) method, which learns the feature embedding and clustering at the same time, and reported state-of-the-art face clustering performances on several commonly used face image databases [46]. In this section, to further illustrate the benefits of the proposed GDM, particularly the practicality and interpretability of the derived affiliated features, we showcase an intriguing example where feature groups describe important regions that discriminate between faces and flowers.

The experiment comprises of two core steps: Firstly, the feature groups of face region are identified using a "face vs. non-face" binary scheme. Here, we adopt the `Yale Face Database B`[13] for the face dataset and the `17 category flower dataset`[14] as the non-face dataset. Next, the SEC is conducted on the selected feature groups for further evaluation, in comparison to the original full pixels. The selected features (pixels) or feature groups are shown in Fig. 9 (i.e., selected important face region), which look like pencil sketches of portraits upon aggregation. Subsequent clustering on the selected features as a multi-class classification task leads to the results given in Table 5, where significant improvements in the clustering performance in terms of both clustering accuracy and mutual information are reported. Notably, the running time to perform clustering is also significantly reduced when the feature groups are used over the original pixels as feature descriptors.

12. http://www.csie.ntu.edu.tw/~cjlin/libsvm
13. http://www.robots.ox.ac.uk/~vgg/data/flowers
14. http://cvc.yale.edu/projects/yalefacesB/yalefacesB.html

6.2 Conclusion

Today, modern databases with big dimensionality are experiencing a growing trend. State-of-the-art approaches that require the calculations of pairwise feature correlations in their algorithmic designs have not coped well on such database, since the computation burden of m^2 is often impractical. In this paper, our observations from several real-world databases have established that an extremely small portion of the feature pairs contribute significantly to the underlying feature interactions, i.e., there is a presence of sparse correlations, and there exists feature groups that are highly correlated. Taking the cue, we then embarked on a comprehensive study on potential correlated informative features or feature groups using the concepts of support feature and affiliated feature, to fill in the research gap that has been identified. In particular, our proposed GDM embeds an explicit incorporation of both linear and nonlinear correlation measures as constraints in the learning model to filter out large number of non-contributing correlations that could otherwise confuse a classifier, while identifying the correlated and informative feature groups. Notably, the affiliated features are constructed in the proposed method without any additional cost, since they are generated along with the support features. Besides, we also demonstrated the proposed method on one-class learning, where notable acceleration can be achieved by GDM-OC from big

TABLE 5
Clustering Performance Results Using Original and Selected Pixels for Face Recognition from 5 Runs (i.e., the Value Is in the Form of mean $\pm$ std.)

Features Used	Original	Support-Affiliated Feature Groups
Clustering Accu. (in percent)	32.67 ± 0.20	$\mathbf{34.98} \pm 0.36$
Mutual Info. (in percent)	45.36 ± 0.18	$\mathbf{47.22} \pm 0.29$
Elapsed Time (in sec.)	320.15 ± 1.29	$\mathbf{87.49} \pm \mathbf{1.07}$

dimensional one-class data. Further, to identify robust informative features with minimal sampling bias, ECV is designed to embed the cross validation scheme in seeking for stable and robust feature sets that are consistent across different data folds. Extensive empirical studies have been reported on both synthetic and several real-world datasets to reveal the superior prediction accuracy of GDM through the identified SFs, while some feature redundancies via the identification of AFs are useful for enhanced interpretation of the learning tasks. Last but not least, both the theoretical analysis and empirical studies verified the high efficacies of the proposed methodology, which makes trillion correlations feasible when dealing with big dimensional data.

Acknowledgments

This work was conducted within the Rolls-Royce@NTU Corporate Lab with support from the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme. Further, Dr. Ivor W. Tsang is grateful for the support from the ARC Future Fellowship FT130100746 and ARC grant LP150100671.

References

[1] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.

[2] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[3] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. Int. Conf. Mach. Learning*, 2003, pp. 856–863.

[4] Y. Zhai, M. Tan, I. Tsang, and Y. S. Ong, "Discovering support and affiliated features from very high dimensions," in *Proc. Int. Conf. Mach. Learning*, Edinburgh, U.K., Jul. 2012, pp. 1455–1462.

[5] S. Kim and E. P. Xing, "Statistical estimation of correlated genome associations to a quantitative trait network," *PLoS Genet.*, vol. 5, no. 8, p. e1000587, 2009.

[6] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, "Feature grouping and selection over an undirected graph," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, Aug. 2012, pp. 922–930.

[7] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," Arizona State Univ., Phoenix, AZ, USA, 2011, http://jiangtanghu.com/docs/en/FeatureSelection/featureselection_techreport.pdf

[8] Y. Zhai, Y.-S. Ong, and I. W. Tsang, "The emerging 'big dimensionality'," *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 14–26, Aug. 2014.

[9] L. Feng, Y.-S. Ong, M.-H. Lim, and I. W. Tsang, "Memetic search with interdomain learning: A realization between CVRP and CARP," *IEEE Trans. Evol. Comput.*, vol. 19, no. 5, pp. 644–658, Oct. 2015.

[10] A. Gupta, Y.-S. Ong, and L. Feng, "Multifactorial evolution: Towards evolutionary multitasking," *IEEE Trans. Evol. Comput.*, 2015, Doi: 10.1109/TEVC.2015.2458037.

[11] A. J. Brookes, "The essence of SNPs," *Gene*, vol. 234, no. 2, pp. 177–186, Jul. 1999.

[12] S. Orrù, et al., "Psoriasis is associated with a SNP haplotype of the corneodesmosin gene (CDSN)," *Tissue Antigens*, vol. 60, no. 4, pp. 292–298, 2002.

[13] R. Paturi, S. Rajasekaran, and J. Reif, "The light bulb problem," *Inf. Comput.*, vol. 117, no. 2, pp. 187–192, 1995.

[14] P. Achlioptas, B. Schölkopf, and K. Borgwardt, "Two-locus association mapping in subquadratic time," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 726–734.

[15] G. Valiant, "Finding correlations in subquadratic time, with applications to learning parities and juntas," in *Proc. IEEE 53rd Annu. Symp. Foundations Comput. Sci.*, 2012, pp. 11–20.

[16] K. Pearson, *The Life, Letters and Labours of Francis Galton*, (3 vols. in 4 parts). Cambridge, U.K.: Cambridge Univ. Press, 1914–1930.

[17] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[18] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Chambery, France, 1993, pp. 1022–1027.

[19] W. H. Press, et al., *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, Feb. 1993.

[20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[21] M. Omidvar, et al., "Cooperative co-evolution with differential grouping for large scale optimization," *IEEE Trans. Evol. Comput.*, vol. 18, no. 3, pp. 378–393, Jun. 2014.

[22] I. Guyon, *Practical Feature Selection: From Correlation to Causality*. Amsterdam, The Netherlands: IOS Press, 2008.

[23] T. Joachims, "Training linear SVMs in linear time," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Philadelphia, PA, USA, Aug. 2006, pp. 217–226.

[24] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou, "Tighter and convex maximum margin clustering," in *Proc. Int. Conf. Artif. Intell. Statist.*, Clearwater Beach, FL, USA, Apr. 2009, vol. 5, pp. 344–351.

[25] A. Mutapcic and S. Boyd, "Cutting-Set methods for robust convex optimization with pessimizing oracles," *Optim. Methods Softw.*, vol. 24, no. 3, pp. 381–406, 2009.

[26] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, 2008.

[27] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, Jan. 2004.

[28] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Mach. Learn. Res.*, vol. 10, pp. 2013–2038, Dec. 2009.

[29] B. Calderhead and M. Girolami, "Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods," *Interface Focus*, vol. 1, no. 6, pp. 821–835, 2011.

[30] T. Speed, "A correlation for the 21st Century," *Science*, vol. 334, no. 6062, pp. 1502–1503, 2011.

[31] I. W. Tsang, J. T. Kwok, and P.-M. Cheung., "Core vector machines: Fast SVM training on very large data sets," *J. Mach. Learn. Res.*, vol. 6, pp. 363–392, 2005.

[32] H. Wang, et al., "Defragging subgraph features for graph classification," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manag.*, New York, NY, USA, 2015, pp. 1687–1690.

[33] M. Chen, et al., "A unified feature selection framework for graph embedding on high dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1465–1477, Jun. 2015.

[34] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Statist. Soc., Ser. B*, vol. 67, pp. 301–320, 2005.

[35] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, 2008.

[36] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, vol. 64, pp. 115–123, Mar. 2008.

[37] L. W. Zhong and J. T. Kwok, "Efficient sparse modeling with automatic feature grouping," presented at the *Int. Conf. Mach. Learning*, Bellevue, WA, USA, 2011.

[38] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 853–858, May 2010.

[39] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, "An improved GLMNET for L1-regularized logistic regression and support vector machines," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 33–41.

[40] R. P. Nair, et al., "Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways," *Net. Genet.*, vol. 41, no. 2, pp. 199–204, Jan. 2009.

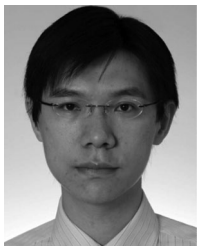
[41] Z. Zhao, et al., "On similarity preserving feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619–632, Mar. 2013.

[42] Y. Saeys, et al., "Robust feature selection using ensemble feature selection techniques," in *Proc. Eur. Conf. Mach. Learning Knowl. Discovery Databases*, 2008, pp. 313–325.

[43] Y. Saeys, et al., "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[44] A. Woznica, P. Nguyen, and A. Kalousis, "Model mining for robust feature selection," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, 2012, pp. 913–921.
[45] B. Schölkopf, et al., "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[46] F. Nie, et al., "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 11, pp. 1796–1808, Nov. 2011.

Yiteng Zhai is currently working toward the PhD degree at the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore. His research interests include feature selection and transfer learning in machine learning area. He was appointed as the coach of the ACM International Collegiate Programming Contest World Finals for NTU team in 2013. He also received two consecutive National Scholarship awards from MOE of the People's Republic of China in 2007 and 2008.

Yew-Soon Ong is a professor of computer science at the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore. He is a principal investigator of the Data Analytics & Complex Systems Programme in the NTU-Rolls Royce Corporate and also the director of the SIMTech-NTU Joint Laboratory on Complex Systems. His research interests include computational intelligence spans across memetic computing, evolutionary design, machine learning, agent-based systems, pattern analysis, and machine intelligence. He is the founding technical editor-in-chief of the *Memetic Computing Journal*, and associate editor of the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Evolutionary Computation*, *IEEE Computational Intelligence Magazine*, *IEEE Transactions on Cybernetics*, *IEEE Transactions on Big Data*, and others. He has published more than 170 refereed articles and delivered technical talks on computational intelligence as keynote, plenary, or invited speaker at international conferences and research institutions worldwide. He received the IEEE Computational Intelligence Magazine Outstanding Paper Award in 2015 and the IEEE Transactions on Evolutionary Computation Outstanding Paper Award in 2012 for his research works in memetic computation. He is a senior member of the IEEE.



Ivor W. Tsang received the PhD degree in computer science from the Hong Kong University of Science and Technology in 2007. He is an Australian future fellow and associate professor with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney (UTS). Before joining UTS, he was the deputy director of the Centre for Computational Intelligence, Nanyang Technological University, Singapore. He has more than 100 research papers published in refereed international journals and conference proceedings, including *Journal of Machine Learning Research*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, Annual Conference on Neural Information Processing Systems, International Conference on Machine Learning, International Conference on Computer Vision, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), etc. In 2009, he was conferred the 2008 Natural Science Award (Class II) by Ministry of Education, China, which recognized his contributions to kernel methods. In 2013, he received his prestigious Australian Research Council Future Fellowship for his research regarding machine learning on big data. Besides these, he had received the prestigious IEEE Transactions on Neural Networks Outstanding 2004 Paper Award in 2006, the 2014 IEEE Transactions on Multimedia Prized Paper Award, and a number of best paper awards and honors from reputable international conferences, including the Best Student Paper Award at CVPR 2010 and the Best Paper Award at ICTAI 2011, etc. He was also awarded the ECCV 2012 Outstanding Reviewer Award.