

Scalable and Accurate Online Feature Selection for Big Data*

KUI YU, Simon Fraser University

XINDONG WU(✉), Hefei University of Technology and University of Vermont

WEI DING, University of Massachusetts Boston

JIAN PEI, Simon Fraser University

Feature selection is important in many big data applications. Two critical challenges closely associate with big data. Firstly, in many big data applications, the dimensionality is extremely high, in millions, and keeps growing. Secondly, big data applications call for highly scalable feature selection algorithms in an online manner such that each feature can be processed in a sequential scan. We present SAOLA, a Scalable and Accurate OnLine Approach for feature selection in this paper. With a theoretical analysis on bounds of the pairwise correlations between features, SAOLA employs novel pairwise comparison techniques and maintain a parsimonious model over time in an online manner. Furthermore, to deal with upcoming features that arrive by groups, we extend the SAOLA algorithm, and then propose a new group-SAOLA algorithm for online group feature selection. The group-SAOLA algorithm can online maintain a set of feature groups that is sparse at the levels of both groups and individual features simultaneously. An empirical study using a series of benchmark real data sets shows that our two algorithms, SAOLA and group-SAOLA, are scalable on data sets of extremely high dimensionality, and have superior performance over the state-of-the-art feature selection methods.

Categories and Subject Descriptors: I.5.2 [Computing Methodologies]: Design Methodology

General Terms: Classifier Design and Evaluation, Feature Evaluation and Selection

Additional Key Words and Phrases: Online feature selection, Extremely high dimensionality, Group features, Big data

1. INTRODUCTION

In data mining and machine learning, the task of feature selection is to choose a subset of relevant features and remove irrelevant and redundant features from high-dimensional data towards maintaining a parsimonious model [Guyon and Elisseeff 2003; Liu and Yu 2005; Xiao et al. 2015; Zhang et al. 2015]. In the era of big data today, many emerging applications, such as social media services, high resolution images, genomic data analysis, and document data analysis, consume data of extremely high dimensionality, in the order of millions or more [Wu et al. 2014; Zhai et al. 2014; Chen et al. 2014; Yu et al. 2015a; Yu et al. 2015b]. For example, the Web Spam Corpus

*A preliminary version of this manuscript with the title “Towards Scalable and Accurate Online Feature Selection for Big Data” was published by the 14th IEEE International Conference on Data Mining (ICDM '14), pp.660-669.

Author's addresses: K. Yu, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, kuiy@sfu.ca; X. Wu (✉ corresponding author), Department of Computer Science, Hefei University of Technology, Hefei, 230009, China and Department of Computer Science, University of Vermont, Burlington, 05405, USA, email: xwu@cs.uvm.edu; W. Ding, Department of Computer Science, University of Massachusetts Boston, Boston, MA, USA, ding@cs.umb.edu; J. Pei, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada, jpei@cs.sfu.ca.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© YYYY ACM 1539-9087/YYYY/01-ARTA \$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

2011 [Wang et al. 2012] collected approximately 16 million features (attributes) for web spam page detection, and the data set from KDD CUP 2010 about using educational data mining to accurately predict student performance includes more than 29 million features. The scalability of feature selection methods becomes critical to tackle millions of features [Zhai et al. 2014].

Moreover, in many applications, feature selection has to be conducted in an online manner. For example, in SINA Weibo, hot topics in behavior in Weibo keep changing daily. When a novel hot topic appears, it may come with a set of new keywords (a.k.a. a set of features). And then some of the new keywords may serve as key features to identify new hot topics. Another example is feature selection in bioinformatics, where acquiring the full set of features for every training instance is expensive because of the high cost in conducting wet lab experiments [Wang et al. 2013]. When it is impossible to wait for a complete set of features, it is practical to conduct feature selection from the features available so far, and consume new features in an online manner as they become available.

To search for a minimal subset of features that leads to the most accurate prediction model, two types of feature selection approaches were proposed in the literature, namely, batch methods [Brown et al. 2012; Woznica et al. 2012; Javed et al. 2014] and online methods [Wu et al. 2013; Wang et al. 2013]. A batch method requires loading the entire training data set into memory. This is obviously not scalable when handling large-scale data sets that exceed memory capability. Moreover, a batch method has to access the full feature set prior to the learning task [Wu et al. 2013; Wang et al. 2013].

Online feature selection has two major approaches. One assumes that the number of features on training data is fixed while the number of data points changes over time, such as the OFS algorithm [Hoi et al. 2012; Wang et al. 2013] that performs feature selection upon each data instance. Different from OFS, the other online method assumes that the number of data instances is fixed while the number of features changes over time, such as the Fast-OSFS [Wu et al. 2013] and alpha-investing algorithms [Zhou et al. 2006]. This approach maintains a best feature subset from the features seen so far by processing each feature upon its arrival. Wang et al. [Wang et al. 2013] further proposed the OGFS (Online Group Feature Selection) algorithm by assuming that feature groups are processed in a sequential scan. It is still an open research problem to efficiently reduce computational cost when the dimensionality is in the scale of millions or more [Wu et al. 2013; Zhai et al. 2014].

In this paper, we propose online feature selection to tackle extremely high-dimensional data for big data analytics, our contributions are as follows.

- We conduct a theoretical analysis to derive a low bound on pairwise correlations between features to effectively and efficiently filter out redundant features.
- With this theoretical analysis, we develop SAOLA, a Scalable and Accurate OnLine Approach for feature selection. The SAOLA algorithm employs novel online pairwise comparisons to maintain a parsimonious model over time. We analyze the upper bound of the gap of information gain between selected features and optimal features.
- To deal with new features that arrive by groups, we extend the SAOLA algorithm, namely, the group-SAOLA algorithm. The group-SAOLA algorithm can online yield a set of feature groups that is sparse between groups as well as within each group for maximizing its predictive performance for classification.
- An extensive empirical study using a series of benchmark data sets illustrates that our two methods, both SAOLA and group-SAOLA, are scalable on data sets of extremely high dimensionality, and have superior performance over the state-of-the-art online feature selection methods.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 proposes our SAOLA algorithm, and Section 4 presents the group-SAOLA algorithm. Section 5 reports our experimental results. Finally, Section 6 concludes the paper and our future work.

2. RELATED WORK

Given a set of input features on a training data set, the problem of feature selection is to select a subset of relevant features from input features without performance degradation of prediction models. There are two types of feature selection approaches proposed in the literature, namely, the batch methods and the online methods.

Standard batch methods can be broadly classified into three categories: filter, wrapper and embedded methods. A wrapper method performs a forward or backward strategy in the space of all possible feature subsets, using a classifier of choice to evaluate each subset. Although this method has high accuracy, the exponential number of possible subsets makes the method computationally expensive in general [Kohavi and John 1997]. The embedded methods attempt to simultaneously maximize classification performance and minimize the number of features used based on a classification or regression model with specific penalties on coefficients of features [Tibshirani 1996; Weston et al. 2000; Zhou et al. 2011].

A filter method is independent of any classifiers, and applies evaluation measures such as distance, information, dependency, or consistency to select features [Dash and Liu 2003; Forman 2003; Peng et al. 2005; Song et al. 2012; Liu et al. 2014]. Then the filter methods build a classifier using selected features. Due to their simplicity and low computational cost, many filter methods have been proposed to solve the feature selection problem, such as the well-established mRMR (minimal-Redundancy-Maximal-Relevance) algorithm [Peng et al. 2005] and the FCBF (Fast Correlation-Based Filter) algorithm [Yu and Liu 2004]. Recently, Zhao et al. [Zhao et al. 2013] proposed a novel framework to consolidate different criteria to handle feature redundancies. To bring almost two decades of research on heuristic filter criteria under a single theoretical interpretation, Brown et al. [Brown et al. 2012] presented a unifying framework for information theoretic feature selection using an optimized loss function of the conditional likelihood of the training labels.

To deal with high dimensionality, Tan et al. [Tan et al. 2010; Tan et al. 2014] proposed the efficient embedded algorithm, the FGM (Feature Generating Machine) algorithm, and Zhai et al. [Zhai et al. 2012] further presented the GDM (Group Discovery Machine) algorithm that outperforms the FGM algorithm.

Group feature selection is also an interesting topic in batch methods, and it selects predictive feature groups rather than individual features. For instance, in image processing, each image can be represented by multiple kinds of groups, such as SIFT for shape information and Color Moment for color information. The lasso method (Least Absolute Shrinkage and Selection Operator) was proposed by Tibshirani [Tibshirani 1996] for shrinkage and variable selection, which minimizes the sum of squared errors with the L_1 penalty on the sum of the absolute values of the coefficients of features. Based on the lasso method, Yuan and Lin [Yuan and Lin 2006] proposed a group lasso method to select grouped variables for accurate prediction in regression. Later, the sparse group lasso criterion as an extension of the group lasso method, was proposed by Friedman et al. [Friedman et al. 2010], which enables to encourage sparsity at the levels of both features and groups simultaneously.

It is difficult for the standard batch method to operate on high dimensional data analytics that calls for dynamic feature selection, because the method has to access all features before feature selection can start.

Online feature selection has two research lines. One assumes that the number of features on training data is fixed while the number of data points changes over time [Hoi et al. 2012]. Recently, Wang et al. [Wang et al. 2013] proposed an online feature selection method, OFS, which assumes data instances are sequentially added.

Different from OFS, the other online approach assumes that the number of data instances is fixed while the number of features changes over time. Perkins and Theiler [Perkins and Theiler 2003] firstly proposed the Grafting algorithm based on a stagewise gradient descent approach. Grafting treats the selection of suitable features as an integral part of learning a predictor in a regularized learning framework, and operates in an incremental iterative fashion, gradually building up a feature set while training a predictor model using gradient descent. Zhou et al. [Zhou et al. 2006] presented Alpha-investing which sequentially considers new features as additions to a predictive model by modeling the candidate feature set as a dynamically generated stream. However, Alpha-investing requires the prior information of the original feature set and never evaluates the redundancy among the selected features.

To tackle the drawbacks, Wu et al. [Wu et al. 2010; Wu et al. 2013] presented the OSFS (Online Streaming Feature Selection) algorithm and its faster version, the Fast-OSFS algorithm. To handle online feature selection with grouped features, Wang et al. [Wang et al. 2013] proposed the OGFS (Online Group Feature Selection) algorithm. However, the computational cost inherent in those three algorithms may still be very expensive or prohibitive when the dimensionality is extremely high in the scale of millions or more.

The big data challenges on efficient online processing and scalability motivate us to develop a scalable and online processing method to deal with data with extremely high dimensionality.

3. THE SAOLA ALGORITHM FOR ONLINE FEATURE SELECTION

3.1. Problem Definition

In general, a training data set D is defined by $D = \{(d_i, c_i), 1 \leq i \leq N\}$, where N is the number of data instances, d_i is a multidimensional vector that contains $numP$ features, and c_i is a class label within the vector of the class attribute C . The feature set F on D is defined by $F = \{F_1, F_2, \dots, F_{numP}\}$. The problem of feature selection on D is to select a subset of relevant features from F to maximize the performance of prediction models. The features in F are categorized into four disjoint groups, namely, strongly relevant, redundant, non-redundant, and irrelevant features [Kohavi and John 1997; Yu and Liu 2004], and the goal of feature selection is to remove redundant or irrelevant features from F while keeping strongly relevant or non-redundant features. The mathematical notations used in this paper are summarized in Table I.

Definition 3.1 (Irrelevant Feature). [Kohavi and John 1997] F_i is an irrelevant feature to C , if and only if $\forall S \subseteq F \setminus \{F_i\}$ and $\forall f_i, \forall c_i, \forall s$ for which $P(S = s, F_i = f_i) > 0$ and $P(C = c_i | S = s, F_i = f_i) = P(C = c_i | S = s)$. ■

Definition 3.2 (Markov Blanket [Koller and Sahami 1995]). A Markov blanket of feature F_i , denoted as $M \subseteq F \setminus \{F_i\}$ makes all other features independent of F_i given M , that is, $\forall Y \in F \setminus (M \cup \{F_i\})$ s.t. $P(F_i | M, Y) = P(F_i | M)$. ■

By Definition 3.2, a redundant feature is defined by [Yu and Liu 2004] as follows.

Definition 3.3 (Redundant Feature). A feature $F_i \in F$ is a redundant feature and hence should be removed from F , if it has a Markov blanket within F . ■

Table I. Summary on mathematical notations

Notation	Mathematical meanings
D	training data set
F	input feature set on D
C	the class attribute
N	the number of data instances
$numP$	the number of features
G	the set of feature groups
L	the conditional likelihood
ℓ	the conditional log-likelihood
G_i	a group of features
Ψ_{t_i}	the set of selected groups at time t_i
S, M, S', ζ	feature subsets within F
X, Y, Z	a single feature ($Y \in F, X \in F, Z \in F$)
x_i, y_i, z_i, c_i	an assignment of values for X, Y, Z , and C
s	an assignment of a set of values of S
d_i	a numP-dimensional data instance
F_i	a N-dimensional feature vector
f_i	a data value of F_i
$S_{(\cdot)}^*$	$S_{(\cdot)}^*$ denote the feature subset selected at time t_i
$ \cdot $	$ \cdot $ returns the size of $S_{(\cdot)}^*$
$P(\cdot \cdot)$	$P(\cdot \cdot)$ denotes the posterior probability of C conditioned on S
δ	relevance threshold
ι	the limitation of the maximum subset size
α	significant level for Fisher's Z-test
ρ	p-value

We also denote D by $D = \{F_i, C\}, 1 \leq i \leq numP$, which is a sequence of features that is presented in a sequential order, where $F_i = \{f_1, f_2, \dots, f_N\}^T$ denotes the i^{th} feature containing N data instances, and C denotes the vector of the class attribute.

If D is processed in a sequential scan, that is, one dimension at a time, we can process high-dimensional data not only with limited memory, but also without requiring its complete set of features available. The challenge is that, as we process one dimension at a time, at any time t_i , how to online maintain a minimum size of feature subset $S_{t_i}^*$ of maximizing its predictive performance for classification. Assuming $S \subseteq F$ is the feature set containing all features available till time t_{i-1} , $S_{t_{i-1}}^*$ represents the selected feature set at t_{i-1} , and F_i is a new coming feature at time t_i , our problem can be formulated as follows:

$$S_{t_i}^* = \arg \min_{S'} \{|S'| : S' = \arg \max_{\zeta \subseteq \{S_{t_{i-1}}^* \cup F_i\}} P(C|\zeta)\}. \quad (1)$$

We can further decompose it into the following key steps:

— Determine the relevance of F_i to C . Firstly, we determine whether Eq.(2) holds or not.

$$P(C|F_i) = P(C). \quad (2)$$

If so, F_i cannot add any additional discriminative power with respect to C , thus F_i should be discarded. Hence, Eq.(2) indicates a new feature F_i that is completely useless by itself with respect to C . If Eq.(2) does not hold, we further evaluate whether F_i carries additional predictive information to C given $S_{t_{i-1}}^*$, that is, whether Eq.(3) below holds. If Eq.(3) holds, F_i has a Markov blanket in $S_{t_{i-1}}^*$, and thus F_i should still be discarded.

$$P(C|S_{t_{i-1}}^*, F_i) = P(C|S_{t_{i-1}}^*). \quad (3)$$

- Calculate $S_{t_i}^*$ with the inclusion of F_i . Once F_i is added to $S_{t_{i-1}}^*$, at time t_i , $S_{t_i} = \{S_{t_{i-1}}^*, F_i\}$, we then solve Eq.(4) to prune S_{t_i} to satisfy Eq.(1).

$$S_{t_i}^* = \arg \max_{\zeta \subseteq S_{t_i}} P(C|\zeta). \quad (4)$$

Accordingly, solving Eq.(1) is decomposed to how to sequentially solve Eq.(2) to Eq.(4) at each time point. Essentially, Eq.(3) and Eq.(4) deal with the problem of feature redundancy.

3.2. Using Mutual Information to Solve Eq.(1)

To solve Eq.(1), we will employ mutual information to calculate correlations between features. Given two features Y and Z , the mutual information between Y and Z is defined as follows.

$$I(Y; Z) = H(Y) - H(Y|Z). \quad (5)$$

The entropy of feature Y is defined as

$$H(Y) = -\sum_{y_i \in Y} P(y_i) \log_2 P(y_i). \quad (6)$$

And the entropy of Y after observing values of another feature Z is defined as

$$H(Y|Z) = -\sum_{z_j \in Z} P(z_j) \sum_{y_i \in Y} P(y_i|z_i) \log_2 P(y_i|z_i), \quad (7)$$

where $P(y_i)$ is the prior probability of value y_i of feature Y , and $P(y_i|z_i)$ is the posterior probability of y_i given the value z_i of feature Z . According to Eq.(6) and Eq.(7), the joint entropy $H(X, Y)$ between features X and Y is defined as follows.

$$\begin{aligned} H(X, Y) &= -\sum_{x_i \in X} \sum_{y_i \in Y} P(x_i, y_i) \log_2 P(x_i, y_i) \\ &= -\sum_{x_i \in X} \sum_{y_i \in Y} P(x_i, y_i) \log_2 P(x_i) P(y_i|x_i) \\ &= -\sum_{x_i \in X} P(x_i) \log_2 P(x_i) - (-\sum_{x_i \in X} \sum_{y_i \in Y} P(x_i, y_i) \log_2 P(y_i|x_i)) \\ &= H(X) + H(Y|X). \end{aligned} \quad (8)$$

From Equations (5) to (8), the conditional mutual information is computed by

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|YZ) \\ &= H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z). \end{aligned} \quad (9)$$

Why can we use mutual information to solve Eq.(1)? Based on the work of [Brown et al. 2012], Eq.(1) is to identify a minimal subset of features to maximize the conditional likelihood of the class attribute C . Let $S = \{S_\theta \cup S_{\bar{\theta}}\}$ represent the feature set containing all features available at time t_i where S_θ indicates the set of selected features and $S_{\bar{\theta}}$ denotes the unselected features. Assuming $p(C|S_\theta)$ denotes the true class distribution while $q(C|S_\theta)$ represents the predicted class distribution given S_θ , then Eq.(1) can be reformulated as $L(C|S_\theta, D) = \prod_{k=1}^N q(c^k|S_\theta^k)$, where $L(C|S_\theta, D)$ denotes the conditional likelihood of the class attribute C given S_θ and D . The conditional log-likelihood of $L(C|S_\theta, D)$ is calculated as follows.

$$\ell(C|S_\theta, D) = \frac{1}{N} \sum_{k=1}^N \log q(c^k|S_\theta^k) \quad (10)$$

By the work of [Brown et al. 2012], Eq.(10) can be re-written as follows.

$$\ell(C|S_\theta, D) = \frac{1}{N} \sum_{k=1}^N \log \frac{q(c^k|S_\theta^k)}{p(c^k|S_\theta^k)} + \frac{1}{N} \sum_{k=1}^N \log \frac{p(c^k|S_\theta^k)}{p(c^k|S^k)} + \frac{1}{N} \sum_{k=1}^N \log p(c^k|S^k) \quad (11)$$

To negate Eq.(11) and use E_{xy} to represent statistical expectation, the following equation holds¹.

$$-\ell(C|S_\theta, D) = E_{xy} \left\{ \log \frac{p(c^k|S_\theta^k)}{q(c^k|S_\theta^k)} \right\} + E_{xy} \left\{ \log \frac{p(c^k|S^k)}{p(c^k|S_\theta^k)} \right\} - E_{xy} \left\{ \log p(c^k|S^k) \right\} \quad (12)$$

In Eq.(12), the first term is a likelihood ratio between the true and predicted class distributions given S_θ , averaged over the input data space. The second term equals to $I(C; S_\theta|S_\theta)$, that is, the conditional mutual information between C and S_θ , given S_θ [Brown et al. 2012]. The final term is $H(C|S)$, the conditional entropy of C given all features, and is an irreducible constant.

Definition 3.4 (Kullback Leibler distance [Kullback and Leibler 1951]). The Kullback Leibler distance between two probability distributions $P(X)$ and $Q(X)$ is defined as $KL(P(X)||Q(X)) = \sum_{x_i \in X} P(x_i) \log \frac{P(x_i)}{Q(x_i)} = E_x \log \left\{ \frac{P(X)}{Q(X)} \right\}$. ■

Then Eq.(12) can be re-written as follows.

$$\lim_{N \rightarrow \infty} -\ell(C|S_\theta, D) = KL(p(C|S_\theta)||q(C|S_\theta)) + I(C; S_\theta|S_\theta) + H(C|S) \quad (13)$$

We estimate the distribution $q(c^k|S_\theta^k)$ using discrete data. The probability of a value c^k of C , $p(C = c^k)$ is estimated by maximum likelihood, the frequency of occurrences of $\{C = c^k\}$ divided by the total number of data instances N . Since the Strong Law of Large Numbers assures that the sample estimate using q converges almost surely to the expected value (the true distribution p), in Eq.(13), $KL(p(C|S_\theta)||q(C|S_\theta))$ will approach zero with a large N [Shlens 2014].

Since $I(C; S) = I(C; S_\theta) + I(C; S_\theta|S_\theta)$ holds, minimizing $I(C; S_\theta|S_\theta)$ is equivalent to maximizing $I(C; S_\theta)$. Accordingly, by Eq.(13), the relationship between the optima of the conditional likelihood and that of the conditional mutual information is achieved as follows.

$$\arg \max_{S_\theta} L(C|S_\theta, D) = \arg \min_{S_\theta} I(C; S_\theta|S_\theta) \quad (14)$$

Eq.(14) concludes that if $I(C; S_\theta|S_\theta)$ is minimal, then $L(C|S_\theta, D)$ is maximal. Therefore, using mutual information as a correlation measure between features, we propose a series of heuristic solutions to Eq.(2), Eq.(3), and Eq.(4) in the next section.

3.3. The Solutions to Equations (2) to (4)

We can apply Definitions 3.2 and 3.3 to solve Eq.(3) and Eq.(4). However, it is computationally expensive to use Definitions 3.2 and 3.3 when the number of features within $S_{t_{i-1}}^*$ is large. Due to evaluating whether F_i is redundant with respect to $S_{t_{i-1}}^*$ using the standard Markov blanket filtering criterion (Definitions 3.2 and 3.3), it is necessary to check all the subsets of $S_{t_{i-1}}^*$ (the total number of subsets is $2^{|S_{t_{i-1}}^*|}$) to determine which subset subsumes the predictive information that F_i has about C , i.e., the Markov blanket of F_i . If such a subset is found, F_i becomes redundant and is removed. When handling a larger number of features, it is computationally prohibitive to check all the subsets of $S_{t_{i-1}}^*$.

Accordingly, methods such as greedy search are a natural to address this problem. In the work of [Wu et al. 2013], a k-greedy search strategy is adopted to evaluate redundant features. It checks all subsets of size less than or equal to ι ($1 \leq \iota \leq |S_{t_{i-1}}^*|$), where ι is a user-defined parameter. However, when the size of $S_{t_{i-1}}^*$ is large, it is still

¹Please refer to Section 3.1 of [Brown et al. 2012] for the details on how to get Eq.(11) and Eq.(12).

computationally prohibitive to evaluate the subsets of size up to ι . Moreover, selecting a proper value of ι is difficult. Therefore, those challenges motivate us to develop a scalable and online processing method to solve Eq.(3) and Eq.(4) for big data analytics.

In this section, to cope with computational complexity, we propose a series of heuristic solutions for Equations (2) to (4) using pairwise comparisons to calculate online the correlations between features, instead of computing the correlations between features conditioned on all feature subsets.

3.3.1. Solving Eq.(2). Assuming $S_{t_{i-1}}^*$ is the selected feature subset at time t_{i-1} , and at time t_i , a new feature F_i comes, to solve Eq.(2), given a relevance threshold δ (we will provide the detailed discussion of the parameter δ in Section 3.4.3), if $I(F_i; C) > \delta$ ($0 \leq \delta < 1$), F_i is said to be a relevant feature to C ; otherwise, F_i is discarded as an irrelevant feature and will never be considered again.

3.3.2. Solving Eq.(3). If F_i is a relevant feature, at time t_i , how can we determine whether F_i should be kept given $S_{t_{i-1}}^*$, that is, whether $I(C; F_i | S_{t_{i-1}}^*) = 0$? If $\exists Y \in S_{t_{i-1}}^*$ such that $I(F_i; C|Y) = 0$, it testifies that adding F_i alone to $S_{t_{i-1}}^*$ does not increase the predictive capability of $S_{t_{i-1}}^*$. With this observation, we solve Eq.(3) with the following lemma.

Lemma 1 $I(X; Y|Z) \geq 0$. ■

Lemma 2 With the current feature subset $S_{t_{i-1}}^*$ at time t_{i-1} and a new feature F_i at time t_i , if $\exists Y \in S_{t_{i-1}}^*$ such that $I(F_i; C|Y) = 0$, then $I(F_i; Y) \geq I(F_i; C)$.

Proof. Considering Eq.(5) and Eq.(9), the following holds.

$$\begin{aligned} I(F_i; C) + I(F_i; Y|C) &= H(F_i) - H(F_i|C) + H(F_i|C) - H(F_i|YC) \\ &= H(F_i) - H(F_i|YC). \end{aligned} \quad (15)$$

$$\begin{aligned} I(F_i; Y) + I(F_i; C|Y) &= H(F_i) - H(F_i|Y) + H(F_i|Y) - H(F_i|YC) \\ &= H(F_i) - H(F_i|YC). \end{aligned} \quad (16)$$

By Equations (15) and (16), the following holds.

$$I(F_i; C|Y) = I(F_i; C) + I(F_i; Y|C) - I(F_i; Y). \quad (17)$$

With Eq.(17), if $I(F_i; C|Y) = 0$ holds, we get the following,

$$I(F_i; Y) = I(F_i; C) + I(F_i; Y|C). \quad (18)$$

Using Eq.(18) and Lemma 1, the bound of $I(F_i; Y)$ is achieved.

$$I(F_i; Y) \geq I(F_i; C). \quad (19)$$

Lemma 2 proposes a pairwise correlation bound between features to testify whether a new feature can increase the predictive capability of the current feature subset. Meanwhile, if $I(F_i; C|Y) = 0$ holds, Lemma 3 answers what the relationship between $I(Y; C)$ and $I(F_i; C)$ is.

Lemma 3 With the current feature subset $S_{t_{i-1}}^*$ at time t_{i-1} and a new feature F_i at time t_i , $\exists Y \in S_{t_{i-1}}^*$, if $I(F_i; C|Y) = 0$ holds, then $I(Y; C) \geq I(F_i; C)$.

Proof. With Eq.(9), we get $I(Y; F_i|C) = I(F_i; Y|C)$. With Eq.(18) and the following equation,

$$I(Y; C|F_i) - I(Y; C) = I(Y; F_i|C) - I(F_i; Y). \quad (20)$$

we get the following,

$$I(Y; C|F_i) = I(Y; C) - I(F_i; C).$$

Case 1: if $I(Y; C|F_i) = 0$, then the following equation holds.

$$I(Y; C) = I(F_i; C). \quad (21)$$

Case 2: if $I(Y; C|F_i) > 0$, we get the following.

$$I(Y; C) > I(F_i; C). \quad (22)$$

By Eq.(21) and Eq.(22), Lemma 3 is proven. ■

According to Lemma 3, we can see that if $I(Y; C|F_i) = 0$ and $I(F_i; C|Y) = 0$, then $I(Y; C)$ exactly equals to $I(F_i; C)$. F_i and Y can replace each other. In Lemma 3, if we only consider Case 2, by Lemma 2, with the current feature subset $S_{t_{i-1}}^*$ at time t_{i-1} and a new feature F_i at time t_i , $\exists Y \in S_{t_{i-1}}^*$, if $I(F_i; C|Y) = 0$ holds, then the following is achieved.

$$I(Y; C) > I(F_i; C) \text{ and } I(F_i; Y) \geq I(F_i; C). \quad (23)$$

With Eq.(23), we deal with Eq.(3) as follows. With a new feature F_i at time t_i , $\exists Y \in S_{t_{i-1}}^*$, if Eq.(23) holds, then F_i is discarded; otherwise, F_i is added to $S_{t_{i-1}}^*$.

3.3.3. Solving Eq.(4). Once F_i is added to $S_{t_{i-1}}^*$ at time t_i , we will check which features within $S_{t_{i-1}}^*$ can be removed due to the new inclusion of F_i . If $\exists Y \in S_{t_{i-1}}^*$ such that $I(C; Y|F_i) = 0$, then Y can be removed from $S_{t_{i-1}}^*$.

Similar to Eq.(17) and Eq.(18), if $I(C; Y|F_i) = 0$, we have $I(Y; F_i) \geq I(Y; C)$. At the same time, if $I(C; Y|F_i) = 0$, similar to Eq.(22), we can get,

$$I(F_i; C) > I(Y; C). \quad (24)$$

With the above analysis, we get the following,

$$I(F_i; C) > I(Y; C) \text{ and } I(Y; F_i) \geq I(Y; C). \quad (25)$$

Accordingly, the solution to Eq.(4) is as follows. With the feature subset $S_{t_i}^*$ at time t_i and $F_i \in S_{t_i}^*$, if $\exists Y \in S_{t_i}^*$ such that Eq.(25) holds, then Y can be removed from $S_{t_i}^*$.

3.4. The SAOLA Algorithm and An Analysis

Using Eq.(23) and Eq.(25), we propose the SAOLA algorithm in detail, as shown in Algorithm 1. The SAOLA algorithm is implemented as follows. At time t_i , as a new feature F_i arrives, if $I(F_i; C) \leq \delta$ holds at Step 5, then F_i is discarded as an irrelevant feature and SAOLA waits for a next coming feature; if not, at Step 11, SAOLA evaluates whether F_i should be kept given the current feature set $S_{t_{i-1}}^*$. If $\exists Y \in S_{t_{i-1}}^*$ such that Eq.(18) holds, we discard F_i and never consider it again. Once F_i is added to $S_{t_{i-1}}^*$ at time t_i , $S_{t_{i-1}}^*$ will be checked whether some features within $S_{t_{i-1}}^*$ can be removed due to the new inclusion of F_i . At Step 16, if $\exists Y \in S_{t_{i-1}}^*$ such that Eq.(20) holds, Y is removed.

3.4.1. The Approximation of SAOLA. To reduce computational cost, the SAOLA algorithm conducts a set of pairwise comparisons between individual features instead of conditioning on a set of features, as the selection criterion for choosing features. This is essentially the idea behind the well-established batch feature selection algorithms, such as mRMR [Peng et al. 2005] and FCBF [Yu and Liu 2004]. Due to pairwise comparisons, our algorithm focuses on finding an approximate Markov blanket (the parents and children of the class attribute in a Bayesian network [Aliferis et al. 2010]) and does not attempt to discover positive interactions between features (there exists a positive interaction between F_i and F_j with respect to C even though F_i is completely useless by itself with respect to C , but F_i can provide significantly discriminative power jointly

ALGORITHM 1: The SAOLA Algorithm.

```

1: Input:  $F_i$ : predictive features,  $C$ : the class attribute;
    $\delta$ : a relevance threshold ( $0 \leq \delta < 1$ ),
    $S_{t_{i-1}}^*$ : the selected feature set at time  $t_{i-1}$ ;
   Output:  $S_{t_i}^*$ : the selected feature set at time  $t_i$ ;
2: repeat
3:   get a new feature  $F_i$  at time  $t_i$ ;
4:   /*Solve Eq.(2)*/
5:   if  $I(F_i; C) \leq \delta$  then
6:     Discard  $F_i$ ;
7:     Go to Step 21;
8:   end if
9:   for each feature  $Y \in S_{t_{i-1}}^*$  do
10:    /*Solve Eq.(3)*/
11:    if  $I(Y; C) > I(F_i; C) \ \& \ I(F_i; Y) \geq I(F_i; C)$  then
12:      Discard  $F_i$ ;
13:      Go to Step 21;
14:    end if
15:    /*Solve Eq.(4)*/
16:    if  $I(F_i; C) > I(Y; C) \ \& \ I(F_i; Y) \geq I(Y; C)$  then
17:       $S_{t_{i-1}}^* = S_{t_{i-1}}^* - Y$ ;
18:    end if
19:  end for
20:   $S_{t_i}^* = S_{t_{i-1}}^* \cup F_i$ ;
21: until no features are available
22: Output  $S_{t_i}^*$ ;

```

with F_j [Jakulin and Bratko 2003; Zhao and Liu 2007]). In the following, we will discuss the upper bound of the gap of information gain between an approximate Markov blanket and an optimal feature set for feature selection.

Given a data set D , by Definition 3.2 in Section 3.1, if we have the optimal feature subset $M \in S$, that is the Markov blanket of C at time t_i , and $S_\theta \in S$ is the feature set selected by SAOLA, and $S_{\bar{\theta}}$ represents $\{S \setminus S_\theta\}$, then according to the chain rule of mutual information, we get $I((S_\theta, S_{\bar{\theta}}); C) = I(S_\theta; C) + I(C; S_{\bar{\theta}}|S_\theta)$. Thus, when S_θ takes the values of the optimal feature subset M , which perfectly captures the underlying distribution $p(C|M)$, then $I(C; S_{\bar{\theta}}|S_\theta)$ would be zero. By Eq.(13), we get the following.

$$\begin{aligned}
-\ell(C|S_\theta, D) &= KL(p(C|S_\theta)||q(C|S_\theta)) + I(C; S_{\bar{\theta}}|S_\theta) + H(C|S) \\
&\leq KL(p(C|S_\theta)||q(C|S_\theta)) + I(C; M) + H(C|S)
\end{aligned} \tag{26}$$

Meanwhile, in Eq.(26), the value of $KL(p(C|S_\theta)||q(C|S_\theta))$ depends on how well q can approximate p , given the selected feature set S_θ . By Eq.(13) in Section 3.2, $KL(p(C|S_\theta)||q(C|S_\theta))$ will approach zero as $N \rightarrow \infty$. Therefore, the upper bound of the gap of information gain between the selected features and optimal features can be re-written as Eq.(27). Closer S_θ is to M , smaller the gap in Eq.(27) is.

$$\lim_{N \rightarrow \infty} -\ell(C|S_\theta, D) \leq I(C; M) + H(C|S) \tag{27}$$

Our empirical results in Section 5.4 have validated that the gap between an optimal algorithm (exact Markov blanket discovery algorithms) and SAOLA, is small using small sample-to-ratio data sets in real-world. Furthermore, SAOLA (pairwise comparisons) is much more scalable than exact Markov blanket discovery algorithms (conditioning on all possible feature subsets) when the number of data instances or dimensionality is large.

3.4.2. Handling Data with Continuous Values. Finally, for data with discrete values, we use the measure of mutual information, while for data with continuous values, we adopt the best known measure of Fisher's Z-test [Peña 2008] to calculate correlations between features. In a Gaussian distribution, $Normal(\mu, \Sigma)$, the population partial correlation $p_{(F_i Y | S)}$ between feature F_i and feature Y given a feature subset S is calculated as follows.

$$p_{(F_i Y | S)} = \frac{-((\sum_{F_i Y S})^{-1})_{F_i Y}}{((\sum_{F_i Y S})^{-1})_{F_i F_i} ((\sum_{F_i Y S})^{-1})_{Y Y}} \quad (28)$$

In Fisher's Z-test, under the null hypothesis of conditional independence between F_i and Y given S , $p_{(F_i Y | S)} = 0$. Assuming α is a given significance level and ρ is the p-value returned by Fisher's Z-test, under the null hypothesis of the conditional independence between F_i and Y , F_i and Y are uncorrelated to each other, if $\rho > \alpha$; otherwise, F_i and Y are correlated to each other, if $\rho \leq \alpha$. Accordingly, at time t , a new feature F_i correlated to C is discarded given S_{t-1}^* , if $\exists Y \in S_{t-1}^*$ s.t. $p_{Y, C} > p_{F_i, C}$ and $p_{Y, F_i} > p_{F_i, C}$.

3.4.3. The Parameters of SAOLA. In Algorithm 1, we discuss the parameters used by the SAOLA algorithm in detail as follows.

- Relevance threshold δ . It is a user-defined parameter to determine relevance thresholds between features and the class attribute. We calculate symmetrical uncertainty [Press et al. 1996] instead of $I(X, Y)$, which is defined by

$$SU(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}.$$

The advantage of $SU(X, Y)$ over $I(X, Y)$ is that $SU(X, Y)$ normalizes the value of $I(X, Y)$ between 0 and 1 to compensate for the bias of $I(X, Y)$ toward features with more values. In general, we set $0 \leq \delta < 1$.

- Correlation bounds of $I(F_i; Y)$. According to Eq.(23) and Eq.(25), at Steps 11 and 16 of the SAOLA algorithm, $I(F_i; C)$ and $I(Y; C)$ ($\min(I(F_i; C), I(Y; C))$) are the correlation bounds of $I(F_i; Y)$, respectively. To further validate the correlation bounds, at Steps 11 and 16, by setting $I(Y; C)$ and $I(F_i; C)$ to $\max(I(F_i; C), I(Y; C))$ respectively, we can derive a variant of the SAOLA algorithm, called SAOLA-max (the SAOLA-max algorithm uses the same parameters as the SAOLA algorithm, except for the correlation bounds of $I(F_i; Y)$ in Steps 11 and 16). We will conduct an empirical study on the SAOLA and SAOLA-max algorithms in Section 5.5.1.
- Selecting a fixed number of features. For different data sets, using the parameters α or δ , SAOLA returns a different number of selected features. Assuming the number of selected features is fixed to k , to modify our SAOLA to select k features, a simple way is to keep the top k features in the current selected feature set $S_{t_i}^*$ with the highest correlations with the class attribute while dropping the other features from $S_{t_i}^*$ after Step 20 in Algorithm 1.

3.4.4. The Time Complexity of SAOLA. The major computation in SAOLA is the computation of the correlations between features (Steps 5 and 11 in Algorithm 1). At time t_i , assuming the total number of features is up to $numP$ and $|S_{t_i}^*|$ is the number of the currently selected feature set, the time complexity of the algorithm is $O(numP |S_{t_i}^*|)$. Accordingly, the time complexity of SAOLA is determined by the number of features within $|S_{t_i}^*|$. But the strategy of online pairwise comparisons guarantees the scalability of SAOLA, even when the size of $|S_{t_i}^*|$ is large.

Comparing to SAOLA, Fast-OSFS employs a k-greedy search strategy to filter out redundant features by checking feature subsets for each feature in $S_{t_i}^*$. At time t_i , the best time complexity of Fast-OSFS is $O(|S_{t_i}^*| \iota^{|S_{t_i}^*|})$, where $\iota^{|S_{t_i}^*|}$ denotes all subsets

of size less than or equal to ι ($1 \leq \iota \leq |S_{t_{i-1}}^*|$) for checking. With respect to Alpha-investing, at time t_i , the time complexity of Alpha-investing is $O(\text{num}P|S_{t_i}^*|^2)$. Since Alpha-investing only considers adding new features but never evaluates the redundancy of selected features, the feature set $S_{t_i}^*$ always has a large size. Thus, when the size of candidate features is extremely high and the size of $|S_{t_i}^*|$ becomes large, Alpha-investing and Fast-OSFS both become computationally intensive or even prohibitive. Moreover, how to select a suitable value of ι for Fast-OSFS in advance is a hard problem, since different data sets may require different ι to search for a best feature subset.

4. A GROUP-SAOLA ALGORITHM FOR ONLINE GROUP FEATURE SELECTION

The SAOLA algorithm selects features only at the individual feature level. When the data possesses certain group structure, the SAOLA algorithm cannot directly deal with features with group structures. In this section, we extend our SAOLA algorithm, and propose a novel group-SAOLA algorithm to select feature groups which are sparse at the levels of both features and groups simultaneously in an online manner.

4.1. Problem Definition

Suppose $G = \{G_1, G_2, \dots, G_i, \dots, G_{\text{num}G}\}$ represents $\text{num}G$ feature groups without overlapping. We denote D by $D = \{G_i, C, 1 \leq i \leq \text{num}G\}$, which is a sequence of feature groups that is presented in a sequential order. If we process those $\text{num}G$ groups in a sequential scan, at any time t_i , the challenge is how to simultaneously optimize selections within each group as well as between those groups to achieve a set of groups, Ψ_{t_i} , containing a set of selected groups that maximizes its predictive performance.

Assuming $G_{i-1} \subset G$ is the set of all feature groups available till time t_{i-1} and G_i is a new coming group at time t_i , our problem can be formulated as follows:

$$\begin{aligned} \Psi_{t_i} = \arg \max_{G_C \subseteq \{G_{i-1} \cup G_i\}} P(C|G_C) \\ \text{s.t.} \\ (a) \forall F_i \in G_j, G_j \subset \Psi_{t_i}, P(C|G_j \setminus \{F_i\}, F_i) \neq P(C|G_j \setminus \{F_i\}) \\ (b) \forall G_j \subset \Psi_{t_i}, P(C|\Psi_{t_i} \setminus G_j, G_j) \neq P(C|\Psi_{t_i} \setminus G_j). \end{aligned} \quad (29)$$

Eq. (29) attempts to yield a solution at time t_i that is sparse at the levels of both intra-groups (constraint (a)) and inter-groups (constraint (b)) simultaneously for maximizing its predictive performance for classification.

Definition 4.1 (Irrelevant groups). If $\exists G_i \subset G$ s.t. $I(C; G_i) = 0$, then G_i is considered as an irrelevant feature group. ■

Definition 4.2 (Group redundancy in inter-groups). If $\exists G_i \subset G$ s.t. $I(C; G_i|G \setminus G_i) = 0$, then G_i is a redundant group. ■

Definition 4.3 (Feature redundancy in intra-groups). $\forall F_i \in G_i$, if $\exists S \subset G_i \setminus \{F_i\}$ s.t. $I(C; F_i|S) = 0$, then F_i can be removed from G_i . ■

With the above definitions, our design to solve Eq. (29) consists of three key steps: at time t_i , firstly, if G_i is an irrelevant group, then we discard it; if not, secondly, we evaluate feature redundancy in G_i to make it as parsimonious as possible at the intra-group level; thirdly, we remove redundant groups from the currently selected groups. The solutions to those three steps are as follows.

— The solution to remove irrelevant groups.

At time t_i , if group $G_i \subset G$ comes, if $\forall F_i \in G_i$ s.t. $I(C; F_i) = 0$ (or given a relevance threshold δ , $I(C; F_i) \leq \delta$ ($0 \leq \delta < 1$)), then G_i is regarded as an irrelevant feature group, and thus it can be discarded.

- The solution to determine group redundancy in inter-groups.
Assume Ψ_{t_i} is the set of groups selected at time t_{i-1} and G_i is a coming group at time t_i . If $\forall F_i \in G_i, \exists F_j \in G_j$, and $G_j \subset \Psi_{t_i}$ s.t. $I(F_j; C) > I(F_i; C)$ and $I(F_j; F_i) \geq I(F_i; C)$, then G_i is a redundant group, and then can be removed.
- The solution to feature redundancy in intra-groups.
If G_i is not a redundant group, we further prune G_i to make it as parsimonious as possible using Theorem 1 in Section 3.2.2. If $\exists F_j \in G_i$ s.t. $\exists Y \in G_i \setminus \{F_j\}, I(Y; C) > I(F_j; C)$ and $I(F_j; Y) \geq I(F_j; C)$ holds, then F_j can be removed from G_i .

4.2. The Group-SAOLA Algorithm

With the above analysis, we propose the Group-SAOLA algorithm in Algorithm 2. In Algorithm 2, from Steps 5 to 8, if G_i is an irrelevant group, it will be discarded. If not, Step 11 and Step 16 prune G_i by removing redundant features from G_i . At Step 22 and Step 26, the group-SAOLA removes both redundant groups in $\{\Psi_{t_{i-1}} \cup G_i\}$ and redundant features in each group currently selected. Our Group-SAOLA algorithm can online yield a set of groups that is sparse between groups as well as within each group simultaneously for maximizing its classification performance at any time t_i .

5. EXPERIMENT RESULTS

5.1. Experiment Setup

We use fourteen benchmark data sets as our test beds, including ten high-dimensional data sets [Aliferis et al. 2010; Yu et al. 2008] and four extremely high-dimensional data sets, as shown in Table II. The first ten high-dimensional data sets include two biomedical data sets (*hiva* and *breast-cancer*), three NIPS 2003 feature selection challenge data sets (*dexter*, *madelon*, and *dorothea*), and two public microarray data sets (*lung-cancer* and *leukemia*), two massive high-dimensional text categorization data sets (*ohsumed* and *apcj-etiology*), and the *thrombin* data set that is chosen from KDD Cup 2001. The last four data sets with extremely high dimensionality are available at the Libsvm data set website².

In the first ten high-dimensional data sets, we use the originally provided training and validation sets for the three NIPS 2003 challenge data sets and the *hiva* data set, and for the remaining six data sets, we randomly select instances for training and testing (see Table II for the number instances for training and testing.). In the *news20* data set, we use the first 9996 data instances for training and the rest for testing while in the *url1* data set, we use the first day data set (*url1*) for training and the second day data set (*url2*) for testing. In the *kdd2010* and *webspam* data sets, we randomly select 20,000 data instances for training, and 100,000 and 78,000 data instances for testing, respectively. Thus, as for the *lung-cancer*, *breast-cancer*, *leukemia*, *ohsumed*, *apcj-etiology*, *thrombin*, *kdd10*, and *webspam* data sets, we run each data set 10 times, and report the highest prediction accuracy and the corresponding running time and number of selected features. Our comparative study compares the SAOLA algorithm with the following algorithms:

- Three state-of-the-art online feature selection methods: Fast-OSFS [Wu et al. 2013], Alpha-investing [Zhou et al. 2006], and OFS [Wang et al. 2013]. Fast-OSFS and Alpha-investing assume features on training data arrive one by one at a time while OFS assumes data examples come one by one;
- Three batch methods: one well-established algorithm of FCBF [Yu and Liu 2004], and two state-of-the-art algorithms, SPSF-LAR [Zhao et al. 2013] and GDM [Zhai et al. 2012].

²<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

ALGORITHM 2: The group-SAOLA Algorithm.

```

1: Input:  $G_i$ : feature group;  $C$ : the class attribute;
    $\delta$ : a relevance threshold ( $0 \leq \delta < 1$ );
    $\Psi_{t_{i-1}}$ : the set of selected groups at time  $t_{i-1}$ ;
   Output:  $\Psi_{t_i}$ : the set of selected groups at time  $t_i$ ;
2: repeat
3:   A new group  $G_i$  comes at time  $t_i$ ;
4:   /*Evaluate irrelevant groups*/
5:   if  $\forall F_i \in G_i, I(F_i; C) \leq \delta$  then
6:     Discard  $G_i$ ;
7:     Go to Step 39;
8:   end if
9:   /*Evaluate feature redundancy in  $G_i$ */
10:  for  $j=1$  to  $|G_i|$  do
11:    if  $\exists Y \in \{G_i - \{F_j\}\}, I(Y; C) > I(F_j; C) \ \& \ I(Y; F_j) \geq I(F_j; C)$  then
12:      Remove  $F_j$  from  $G_i$ ;
13:      Continue;
14:    end if
15:    /*Otherwise*/
16:    if  $I(F_j; C) > I(Y; C) \ \& \ I(F_j; Y) \geq I(Y; C)$  then
17:      Remove  $Y$  from  $G_i$ ;
18:    end if
19:  end for
20:  /*Evaluate group redundancy in  $\{\Psi_{t_{i-1}} \cup G_i\}$ */
21:  for  $j=1$  to  $|\Psi_{t_{i-1}}|$  do
22:    if  $\exists F_k \in G_j \subset \Psi_{t_{i-1}}, \exists F_i \in G_i, I(F_i; C) > I(F_k; C) \ \& \ I(F_i; F_k) \geq I(F_k; C)$  then
23:      Remove  $F_k$  from  $G_j$ ;
24:    end if
25:    /*Otherwise*/
26:    if  $I(F_k; C) > I(F_i; C) \ \& \ I(F_k; F_i) \geq I(F_i; C)$  then
27:      Remove  $F_i$  from  $G_i$ ;
28:    end if
29:    if  $G_j$  is empty then
30:       $\Psi_{t_{i-1}} = \Psi_{t_{i-1}} - G_j$ ;
31:    end if
32:    if  $G_i$  is empty then
33:      Break;
34:    end if
35:  end for
36:  if  $G_i$  is not empty then
37:     $\Psi_{t_i} = \Psi_{t_{i-1}} \cup G_i$ ;
38:  end if
39: until no groups are available
40: Output  $\Psi_{t_i}$ ;

```

The algorithms above are all implemented in MATLAB except for the GDM algorithm that is implemented in C language. We use three classifiers, KNN and J48 provided in the Spider Toolbox³, and SVM⁴ to evaluate a selected feature subset in the experiments. The value of k for the KNN classifier is set to 1 and both SVM and KNN use the linear kernel. All experiments were conducted on a computer with Intel(R) i7-2600, 3.4GHz CPU, and 24GB memory. In the remaining sections, the parameter

³<http://people.kyb.tuebingen.mpg.de/spider/>

⁴<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Table II. The benchmark data sets

Dataset	Number of features	Number of training instances	Number of testing instances
madelon	500	2,000	600
hiva	1,617	3,845	384
leukemia	7,129	48	24
lung-cancer	12,533	121	60
ohsumed	14,373	3,400	1,600
breast-cancer	17,816	190	96
dexter	20,000	300	300
apcj-etiology	28,228	11,000	4,779
dorothea	100,000	800	300
thrombin	139,351	2,000	543
news20	1,355,191	9,996	10,000
url1	3,231,961	20,000	20,000
webspam	16,609,143	20,000	78,000
kdd2010	29,890,095	20,000	100,000

δ for SAOLA is set to 0 for discrete data while the significance level α for SAOLA is set to 0.01 for Fisher’s Z-test for continuous data (the effect of δ and α on SAOLA was given in Section 5.4.3.). The significance level is set to 0.01 for Fast-OSFS, and for Alpha-investing, the parameters are set to the values used in [Zhou et al. 2006].

We evaluate SAOLA and its rivals based on prediction accuracy, error bar, AUC, sizes of selected feature subsets, and running time. In the remaining sections, to further analyze the prediction accuracies and AUC of SAOLA against its rivals, we conduct the following statistical comparisons.

- Paired t-tests are conducted at a 95% significance level and the win/tie/lose (w/t/l for short) counts are summarized.
- To validate whether SAOLA and its rivals have no significant difference in prediction accuracy or AUC, we conduct the Friedman test at a 95% significance level [Demšar 2006], under the null-hypothesis, which states that the performance of SAOLA and that of its rivals have no significant difference, and calculate the average ranks using the Friedman test (how to calculate the average ranks, please see [Demšar 2006]).
- When the null-hypothesis at the Friedman test is rejected, we proceed with the Nemenyi test [Demšar 2006] as a post-hoc test. With the Nemenyi test, the performance of two methods is significantly different if the corresponding average ranks differ by at least the critical difference (how to calculate the critical difference, please see [Demšar 2006]).

We organize the remaining parts as follows. Section 5.2 compares SAOLA with online feature selection algorithms. Section 5.3 gives a comparison of SAOLA with batch feature selection methods, Section 5.4 compares SAOLA with Markov blanket discovery methods, and Section 5.5 conducts an analysis of the effect of parameters on SAOLA. Section 5.6 compares Group-SAOLA with its rivals.

5.2. Comparison of SAOLA with Three Online Algorithms

5.2.1. Prediction Accuracy, Running Time and the Number of Selected Features of SAOLA. Since Fast-OSFS and Alpha-investing can only deal with the first ten high-dimensional data sets in Table II due to high computational cost, we compare them with SAOLA only on the first ten high-dimensional data sets. Accordingly, in the following tables, the notation “-” denotes an algorithm fails to a data set because of excessive running time.

The OFS algorithm is a recently proposed online feature selection method. Since OFS uses a user-defined parameter k to control the size of the final selected feature subset, we set k , i.e., the number of selected features to the top 5, 10, 15, ..., 100 fea-

Table III. Prediction accuracy (J48)

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
dexter	0.8133	0.8200	0.5000	0.5667
lung-cancer	0.9500	0.9000	0.8333	0.8667
hiva	0.9661	0.9635	0.9635	0.9635
breast-cancer	0.7917	0.8854	0.7187	0.8333
leukemia	0.9583	0.9583	0.6667	0.9583
madelon	0.6083	0.6100	0.6067	0.6367
ohsumed	0.9437	0.9450	0.9331	0.9431
apcj-etiology	0.9872	0.9868	0.9828	0.9872
dorothea	0.9343	0.9371	0.9343	0.9371
thrombin	0.9613	0.9595	0.9613	0.9374
news20	0.8276	-	-	0.7332
url1	0.9744	-	-	0.9720
kdd10	0.8723	-	-	0.8577
webspam	0.9611	-	-	0.9689
w/t/l	-	1/8/1	5/5/0	5/7/2

Table IV. Prediction accuracy (KNN)

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
dexter	0.7600	0.7800	0.5000	0.5400
lung-cancer	0.9833	0.9667	0.9167	0.8500
hiva	0.9635	0.9635	0.9531	0.9661
breast-cancer	0.8646	0.8542	0.6875	0.6979
leukemia	0.9167	0.7917	0.6250	0.8750
madelon	0.5617	0.5283	0.5767	0.6433
ohsumed	0.9275	0.9306	0.9325	0.9431
apcj-etiology	0.9793	0.9702	0.9851	0.9872
dorothea	0.9200	0.9457	0.7400	0.9086
thrombin	0.9374	0.9300	0.9371	0.9411
news20	0.7755	-	-	0.6884
url1	0.9627	-	-	0.9607
kdd10	0.8780	-	-	0.7755
webspam	0.9532	-	-	0.9516
w/t/l	-	4/4/2	6/3/1	7/5/2

Table V. Prediction accuracy (SVM)

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
dexter	0.8500	0.8100	0.5000	0.5000
lung-cancer	0.9833	0.9500	0.9167	0.7833
hiva	0.9635	0.9635	0.9635	0.9635
breast-cancer	0.8750	0.8854	0.7188	0.7812
leukemia	0.9583	0.7500	0.6667	0.8333
madelon	0.6217	0.6227	0.6383	0.6117
ohsumed	0.9431	0.9438	0.9431	0.9431
apcj-etiology	0.9872	0.9872	0.9872	0.9872
dorothea	0.9286	0.9371	0.9086	0.9029
thrombin	0.9116	0.9116	0.9153	0.9245
news20	0.8721	-	-	0.4993
url1	0.9645	-	-	0.9681
kdd10	0.8727	-	-	0.8852
webspam	0.9123	-	-	0.8897
w/t/l	-	3/6/1	5/4/1	8/4/2

tures, then selecting the feature set with the highest prediction accuracy as the reporting result.

Tables III, IV, and V summarize the prediction accuracies of SAOLA against Fast-OSFS, Alpha-investing, and OFS using the KNN, J48 and SVM classifiers. The win/tie/loss (w/t/l for short) counts of SAOLA against Fast-OSFS, Alpha-investing, and OFS are summarized in Tables III, IV, and V. The highest prediction accuracy is high-

lighted in bold face. Tables VI and VII give the number of selected features and running time of SAOLA, Fast-OSFS, Alpha-investing, and OFS. We have the following observations.

(1) SAOLA vs. Fast-OSFS. With the counts of w/t/l in Tables III and V, we observe that SAOLA is very competitive with Fast-OSFS. In Table IV, we can see that SAOLA is superior to Fast-OSFS. Fast-OSFS selects fewer features than SAOLA on all data sets as shown in Table VI. The explanation is that Fast-OSFS employs a k-greedy search strategy to filter out redundant features by checking the feature subsets in the current feature set for each feature while SAOLA only uses pairwise comparisons. But as shown in Table VII, this strategy makes Fast-OSFS very expensive in computation and even prohibitive on some data sets, such as *apcj-etiology* and the last four extremely high-dimensional data sets of Table II, as the size of the current feature set is large at each time point.

(2) SAOLA vs. Alpha-investing. From Tables III to V, we can see that SAOLA outperforms Alpha-investing on most data sets using the three classifiers. Alpha-investing selects many more features than SAOLA on *ohsumed*, *apcj-etiology*, *dorothea*, and *thrombin* since Alpha-investing only considers to add new features but never evaluates the redundancy of selected features. An exception is that Alpha-investing only selects one feature on the *dexter* data set. A possible explanation is that the *dexter* data set is a very sparse real-valued data set. Furthermore, Alpha-investing is less efficient than SAOLA as shown in Table VII.

Table VI. Number of selected features

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
dexter	21	9	1	85
lung-cancer	35	6	7	60
hiva	12	5	48	10
breast-cancer	46	7	2	10
leukemia	17	5	2	45
madelon	3	3	4	65
ohsumed	65	11	297	10
apcj-etiology	75	67	634	10
dorothea	63	5	113	60
thrombin	20	9	60	40
news20	212	-	-	85
url1	64	-	-	100
kdd10	180	-	-	90
webspam	51	-	-	85

To validate whether SAOLA, Fast-OSFS, and Alpha-investing have no significant difference in prediction accuracy, with the Friedman test at 95% significance level, under the null-hypothesis, which states that the performance of SAOLA and that of Fast-OSFS and Alpha-investing have no significant difference, with respect to J48 in Table III, the average ranks for SAOLA, Fast-OSFS, and Alpha-investing are 2.35, 2.40, and 1.25 (the higher the average rank, the better the performance), respectively. The null-hypothesis is rejected. Then we proceed with the Nemenyi test as a post-hoc test. With the Nemenyi test, the performance of two methods is significantly different if the corresponding average ranks differ by at least the critical difference. With the Nemenyi test, the critical difference is up to 1.047. Thus, with the critical difference and the average ranks calculated above, the prediction accuracy of SAOLA and that of Fast-OSFS have no significant difference, but SAOLA is significantly better than Alpha-investing.

As for the KNN classifier, the average ranks for SAOLA, Fast-OSFS, and Alpha-investing are 2.45, 1.85, and 1.705 in Table IV, respectively. Meanwhile, as for SVM

in Table V, the average ranks for SAOLA, Fast-OSFS, and Alpha-investing are 2.30, 2.25, and 1.45, respectively. Using KNN and SVM, the null-hypothesis cannot be rejected, and thus, the prediction accuracy of SAOLA and that of Fast-OSFS and Alpha-investing have no significant difference.

In summary, in prediction accuracy, SAOLA is very competitive with Fast-OSFS, and is superior to Alpha-investing. Furthermore, Fast-OSFS and Alpha-investing cannot deal with extremely high-dimensional data sets due to computational cost while SAOLA is accurate and scalable.

Table VII. Running time (seconds)

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
dexter	3	4	6	1
lung-cancer	6	4	2	1
hiva	1	36	7	1
breast-cancer	5	4	3	1
leukemia	2	2	1	0.1
madelon	0.1	0.1	0.1	0.1
ohsumed	6	343	497	9
apcj-etiology	22	> 3 days	9,781	100
dorothea	58	375	457	10
thrombin	63	18,576	291	40
news20	944	-	-	1,572
url1	200	-	-	1,837
kdd10	1,056	-	-	28,536
webspam	1,456	-	-	18,342

(3) SAOLA vs. OFS. With Tables III to V, we evaluate whether the performance of SAOLA and that of OFS have no significant difference in prediction accuracy using the Friedman test at 95% significance level. For the J48 and SVM classifiers, we observe the same average ranks for SAOLA and OFS, 1.64 and 1.36, respectively. Regarding SVM, the average ranks for SAOLA and OFS are 1.61 and 1.39, respectively. Accordingly, although SAOLA is better than OFS on prediction accuracy using the w/t/l counts and the average ranks, SAOLA and OFS have no significant difference in prediction accuracy.

However, from Table VI, we can see that SAOLA selects fewer features than OSF on all data sets except for *hiva*, *breast-cancer*, *ohsumed*, *apcj-etiology*, *news20*, and *kdd10*. Moreover, Table VII gives the running time of SAOLA and OFS. As for OFS, we record the running time of the feature subset with the highest accuracy as its running time. SAOLA is faster than OFS, except for the *dorothea* and *thrombin* data sets. The *dorothea* and *thrombin* data sets only include 800 samples and 2000 samples, respectively. When the number of data samples becomes large and the number of features of training data is increased to millions, OFS become very costly, and SAOLA is still scalable and efficient. The explanation is that the time complexity of SAOLA is determined by the number of features within the currently selected feature set, and the strategy of online pairwise comparisons makes SAOLA very scalable, even when the size of the current feature set is large. Moreover, setting a desirable size of a feature set selected by OFS in advance is a non-trivial task.

5.2.2. AUC and Error Bar of SAOLA. In this section, we further evaluate SAOLA and the three online algorithms using the error bar and AUC metrics.

Table VIII reports the average AUC of J48, KNN and SVM for each algorithm. Using the w/t/l counts, we can see that our SAOLA is better than Fast-OSFS and Alpha-investing. To further validate whether SAOLA, Fast-OSFS, and Alpha-investing have

Table VIII. AUC of SAOLA, Fast-OSFS, and Alpha-investing

Dataset	SAOLA	Fast-OSFS	Alpha-investing
dexter	0.8088	0.8033	0.5000
lung-cancer	0.9407	0.9286	0.8298
hiva	0.5349	0.5229	0.5326
breast-cancer	0.8111	0.7996	0.5303
leukemia	0.9404	0.8811	0.6543
madelon	0.5972	0.5883	0.6072
ohsumed	0.5559	0.6244	0.5503
apcj-etiology	0.5256	0.5250	0.4987
dorothea	0.7297	0.8219	0.6453
thrombin	0.7849	0.7913	0.8026
average rank	2.6	2.1	1.3
w/t/l	-	5/3/2	6/2/2

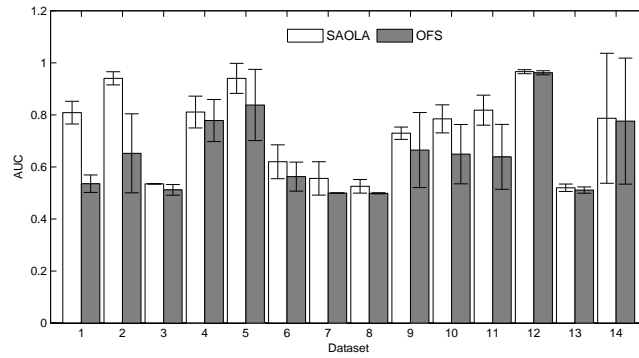


Fig. 1. The AUC of SAOLA and OFS (the labels of the x-axis from 1 to 14 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin; 11. news20; 12. url1; 13. kdd10; 14. webspam)

no significant difference in AUC, with the Friedman test, the null-hypothesis is rejected, and the average ranks calculated for SAOLA, Fast-OSFS, and Alpha-investing are 2.6, 2.1, and 1.3, respectively.

Then we proceed with the Nemenyi test as a post-hoc test. With the Nemenyi test, the performance of two methods is significantly different if the corresponding average ranks differ by at least the critical difference. With the Nemenyi test, the critical difference is up to 1.047. Accordingly, the AUC of SAOLA and that of Fast-OSFS have no significant difference, but SAOLA is significantly better than Alpha-investing.

Figure 1 shows the average AUC of J48, KNN and SVM and its standard deviation for SAOLA and OFS. We can see that SAOLA outperforms to OFS on all 14 data sets.

From Table VIII and Figure 1, we can conclude that none of SAOLA, Fast-OSFS, Alpha-investing, and OFS can effectively deal with highly class-imbalanced data sets, such as *hiva*, *apcj-etiology*, and *ohsumed*.

Finally, we give the error bars of SAOLA, Fast-OSFS, Alpha-investing, and OFS. Since we randomly select instances from the *lung-cancer*, *breast-cancer*, *leukemia*, *ohsumed*, *apcj-etiology*, *thrombin*, *kdd10*, and *webspam* data sets for training and testing, we only report the error bars of the four online algorithms on those eight data sets using the SVM classifier. For each data set, we randomly run each algorithm 10 times, and then compute the average prediction accuracy and the corresponding standard deviation. Figure 2 plots the average prediction accuracy and an error bar that denotes a distance of the standard deviation above and below this

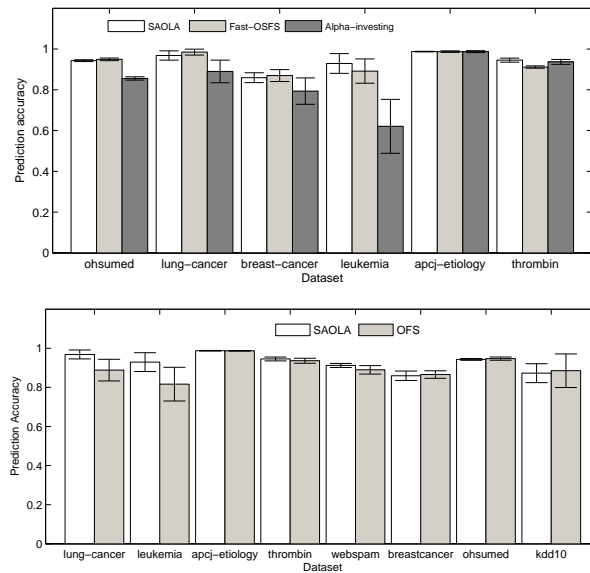


Fig. 2. The error bars of SAOLA, Fast-OSFS, Alpha-investing, and OFS

average prediction accuracy on each data set. Figure 2 shows that SAOLA achieves a higher average prediction accuracy and a lower standard deviation than Alpha-investing and OFS while being highly comparable with Fast-OSFS. This further confirms that SAOLA is very competitive with Fast-OSFS and superior to Alpha-investing and OFS.

5.2.3. Stability of SAOLA. The stability of feature selection is one of the criteria to measure the performance of a feature selection algorithm by quantifying the ‘similarity’ between two selected feature sets, and was first discussed by Kalousis et al. [Kalousis et al. 2007]. In this section, we employ the measure proposed by Yu et al. [Yu et al. 2008] to evaluate the stabilities of SAOLA, Fast-OSFS, Alpha-investing, and OFS. This measure constructs a weighted complete bipartite graph, where the two node sets correspond to two different feature sets, and weights assigned to the arcs are the normalized mutual information between the features at the nodes, also sometimes referred to as the symmetrical uncertainty. The Hungarian algorithm is then applied to identify the maximum weighted matching between the two node sets, and the overall similarity between two sets is the final matching cost.

To evaluate the stabilities, each data set was randomly partitioned into five folds, each fold containing 1/5 of all the samples. SAOLA and its rivals under comparison were repeatedly applied to four out of the five folds. This process was repeated 30 times to generate different subsamples for each data set. Then the average stabilities over 30 subsamples on each data set are as the results of SAOLA, Fast-OSFS, Alpha-investing, and OFS.

Figure 3 shows the stabilities of SAOLA, Fast-OSFS, and Alpha-investing (we do not plot the stability of Alpha-investing on the *dexter* data set, since Alpha-investing only selects one feature on the *dexter* data set.). We can see that the Alpha-investing algorithm is the most stable feature selection algorithm among the three online methods. The explanation is that the Alpha-investing algorithm only considers adding features while never removes redundant features. SAOLA and Fast-OSFS aim to select a minimum subset of features necessary for constructing a classifier of best predictive

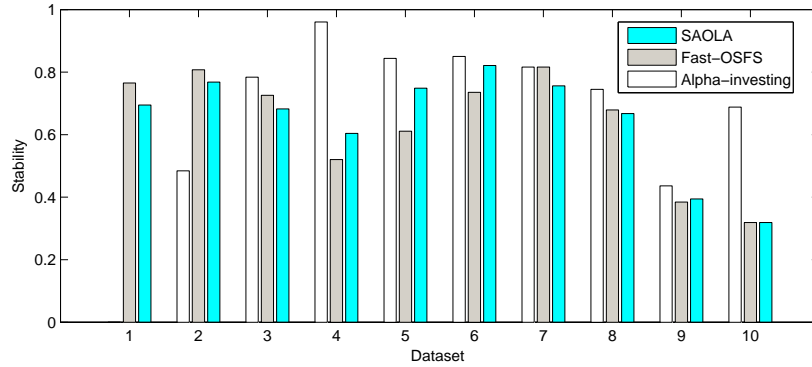


Fig. 3. The stabilities of SAOLA, Fast-OSFS, and Alpha-investing (The labels of the x-axis from 1 to 10 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin)

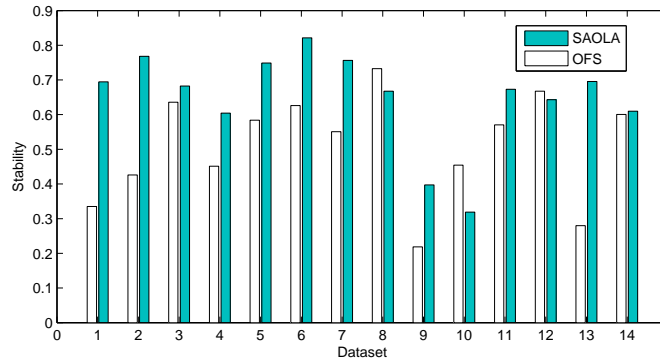


Fig. 4. The stabilities of SAOLA and OFS (The labels of the x-axis from 1 to 14 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin; 11. news20; 12. url; 13. kdd10; 14. webspam)

accuracy and discard features which are relevant to the class attribute but highly correlated to the selected ones. Among a set of highly correlated features, different ones may be selected under different settings of SAOLA and Fast-OSFS. Therefore, from Figure 3, we can see that SAOLA is very competitive with Fast-OSFS.

Meanwhile, from Figure 4, we can observe that SAOLA is more stable than OFS. Such observation illustrates that even if OFS can select large subsets of features, it is still less stable than SAOLA. The possible explanation is that OFS assumes that data examples come one by one while SAOLA assume that features on training data arrive one by one at a time.

5.2.4. The Effect of Large Data Sets on SAOLA. To evaluate the effect of large data sets on SAOLA, we use the data sets, *connect-4* (60,000/7,557 objects and 126 attributes) (note: 60,000/7,557 objects denote 60,000 data instances for training while 7,557 ones for testing), *ijcnn1* (190,000/1,681 objects and 22 attributes), *covtype* (571,012/10,000 objects and 54 attributes), *poker* (1,020,000/5,010 objects and 11 attributes), and *real-*

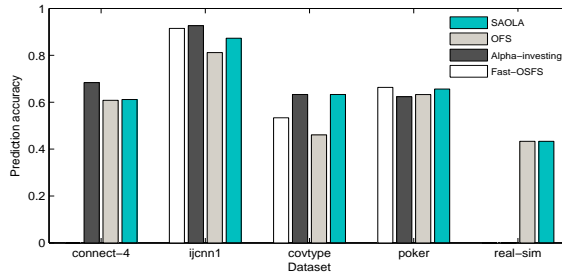


Fig. 5. Prediction accuracy of SAOLA, Fast-OSFS, Alpha-investing, and OFS

sim (70,000/2,309 objects and 20,958 attributes) from the machine learning data set repository ⁵.

Table IX gives the running time of SAOLA and its rivals. In Table IX, “-” denotes that an algorithm fails to the data set due to the high computational cost (exceeding three days). We can see that SAOLA is more scalable to deal with data with large numbers of data instances than Fast-OSFS, Alpha-investing, and OFS. Furthermore, Figure 5 gives the prediction accuracy of the four algorithms using the SVM classifier. Since Fast-OSFS fails on the *connect-4* and *real-sim* data sets while Alpha-investing cannot run on the *real-sim* data set, Figure 5 does not plot the prediction accuracies of Fast-OSFS and Alpha-investing on those data sets. SAOLA is better than Fast-OSFS and OFS and competitive with Alpha-investing on prediction accuracy.

Table IX. Running time of SAOLA and its three rivals (in seconds)

Dataset	SAOLA	Fast-OSFS	Alpha-investing	OFS
connect-4	2	-	38	310
ijcnn1	0.45	2	0.75	2
covtype	13	15,324	48	8,846
poker	0.26	0.42	0.92	10
real-sim	1,223	-	-	1,013

5.3. Comparison with the Three Batch Methods

5.3.1. Running Time, the Number of Selected Features, and Prediction Accuracy of SAOLA. Since FCBF and SPSF-LAR can only deal with the first ten high-dimensional data sets in Table II, in the following tables and figures, we compare FCBF and SPSF-LAR with our proposed algorithm only on those ten high-dimensional data sets in terms of size of selected feature subsets, running time, and prediction accuracy. The information threshold for FCBF is set to 0. We set the user-defined parameter k , i.e., the number of selected features to the top 5, 10, 15, ..., 65 features for the SPSF-LAR algorithm, choose the feature subsets of the highest prediction accuracy, and record the running time and the size of this feature set as the running time and the number of selected features of SPSF-LAR, respectively.

We also select the GDM algorithm [Zhai et al. 2012] which is one of the most recent batch feature selection methods in dealing with very large dimensionality. The GDM algorithm is an efficient embedded feature selection method using cutting plane strategy and correlation measures as constraints to minimize the correlation among the selected features. GDM uses a user-defined parameter to control the size of the final selected feature subset. We set the selected feature subset sizes to the top 10,

⁵<http://mldata.org/repository/data/>

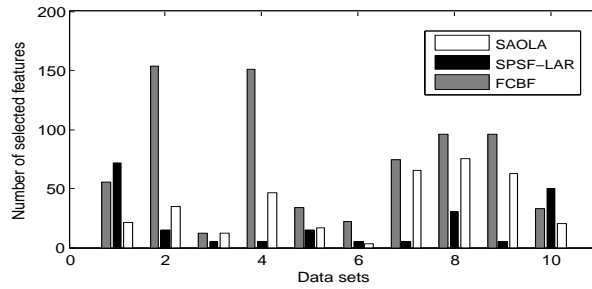


Fig. 6. Number of selected features (The labels of the x-axis from 1 to 10 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin)

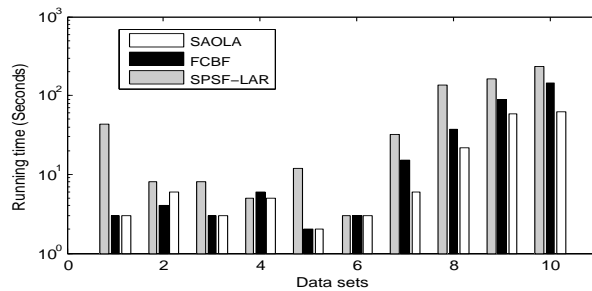


Fig. 7. Running time (the labels of the x-axis are the same as the labels of the x-axis in Figure 6.)

20, 30, ..., 260 features for the GDM algorithm, report the running time of the feature subset with the highest accuracy as the running time of GDM, and choose the highest prediction accuracies achieved among those selected feature subsets.

From Figure 6, we can conclude that FCBF selects the most features among SAOLA, FCBF and SPSF-LAR while SAOLA and SPSF-LAR are similar to each other. As shown in Figure 7, we can observe that SAOLA is the fastest algorithm among SAOLA, FCBF and SPSF-LAR while SPSF-LAR is the slowest. The explanation is that the time complexity of the algorithm is $O(\text{num}P|S_{t_i}^*|)$ where $\text{num}P$ is the number of features and $|S_{t_i}^*|$ is the number of selected features at time t_i , while the time complexity of SPFS-LAR is $O(\text{num}PNk + Nk^3)$ where N is the number of data samples and k the number of selected features.

The computational costs of SAOLA and FCBF are very competitive since both of them employ pairwise comparisons to calculate the correlations between features. But when the number of data instances or the number of features is large, SAOLA is faster than FCBF, such as on *ohsumed*, *apcj-etiology*, *dorothea*, and *thrombin*. A possible reason is that SAOLA online evaluates both the new coming features (Step 11) and the current feature set (Step 16) at each time point to make the selected feature set as parsimonious as possible, since the size of the selected feature set at each time point is the key to determine the running time of SAOLA. With the same information threshold δ , FCBF prefers to selecting more features than SAOLA.

Figure 8 shows the running time of SAOLA against GDM. Since GDM is implemented in C++, we developed a C++ version of SAOLA for the comparison with GDM, in addition to its Matlab version. In Figure 8, we only give the last four data sets with extremely high dimensionality in Table II, since on the the first ten data sets, the running time of both SAOLA and GDM is no more than ten seconds. We can see that

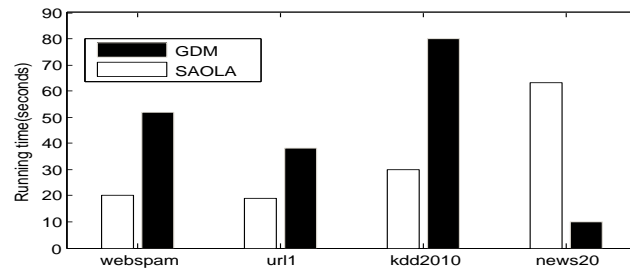


Fig. 8. Running time of SAOLA and GDM

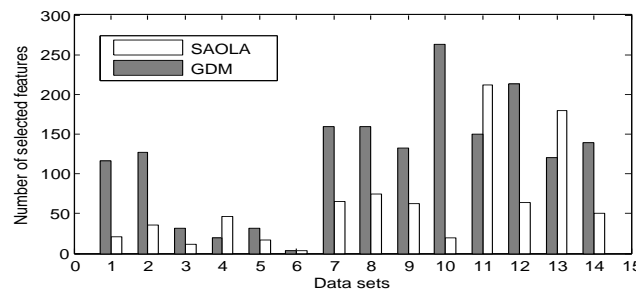


Fig. 9. Number of selected features (The labels of the x-axis from 1 to 14 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin; 11. news20; 12. url1; 13. kdd10; 14. webspam)

although GDM is a wrapper-like feature selection method, both GDM and SAOLA are very efficient to handle extremely high-dimensional data sets. Except for the *news20* data set, SAOLA is a little faster than GDM. On the sparse data sets, SAOLA is faster than GDM, while on the dense data sets, such as the *news20* data set, GDM is faster than SAOLA. Finally, Figure 9 reports the number of selected features of SAOLA comparing to GDM. Except for the *breast-cancer* data set, SAOLA selects fewer features than GDM to achieve the very competitive prediction accuracy with GDM.

Finally, Tables X to XII report the prediction accuracies of SAOLA against FCBF, SPSF-LAR, and GDM. With the counts of w/t/l in the last rows of Tables X to XII, we can see that even without requiring the entire feature set on a training data set in advance, SAOLA is still very competitive with FCBF, SPSF-LAR and GDM in prediction accuracy.

Using the Friedman test at 95% significance level, for J48, the average ranks for SAOLA, FCBF and SPSF-LAR are 2.05, 1.90, and 2.05, respectively. For KNN, the average ranks for SAOLA, Fast-OSFS, and SPSF-LAR are 1.90, 2.25 and 1.85, respectively, while regarding SVM, the average ranks are 2.15, 1.65, and 2.20. Thus, with the Friedman test at 95% significance level, using KNN, J48 and SVM, the null-hypothesis cannot be rejected, and thus SAOLA, FCBF and SPSF-LAR have no significant difference in prediction accuracy. Accordingly, we conclude that the performance of SAOLA is highly comparable to that of FCBF and SPSF-LAR.

With regard to the prediction accuracies of SAOLA and GDM, we can see that our algorithm is very competitive with GDM on J48, KNN, and SVM. With the Friedman test at 95% significance level, for J48, the average ranks for SAOLA and GDM are 1.32 and 1.68, respectively. As for KNN, the average ranks for SAOLA and GDM are 1.39 and 1.61, respectively. Using KNN and J48, the null-hypothesis cannot be rejected,

Table X. Prediction accuracy (J48)

Dataset	SAOLA	FCBF	SPSF-LAR	GDM
dexter	0.8133	0.8567	0.8700	0.9100
lung-cancer	0.9500	0.9500	0.9833	0.9833
hiva	0.9661	0.9661	0.9635	0.9661
breast-cancer	0.7917	0.8125	0.8958	0.4792
leukemia	0.9583	0.9583	0.9583	1.0000
madelon	0.6083	0.6067	0.6183	0.5833
ohsumed	0.9437	0.9444	0.9431	0.9438
apcj-etiology	0.9872	0.9866	0.9872	0.9879
dorothea	0.9343	0.9314	0.9029	0.9371
thrombin	0.9613	0.9576	0.9558	0.7300
news20	0.8276	-	-	0.7354
url1	0.9744	-	-	0.9765
kdd10	0.8723	-	-	0.8779
webspam	0.9611	-	-	0.9617
w/t/l	-	0/8/2	1/5/4	4/7/3

Table XI. Prediction accuracy (KNN)

Dataset	SAOLA	FCBF	SPSF-LAR	GDM
dexter	0.7600	0.7967	0.7233	0.9100
lung-cancer	0.9833	0.9500	0.9833	0.9833
hiva	0.9635	0.9609	0.9635	0.9661
breast-cancer	0.8646	0.8333	0.8229	0.4792
leukemia	0.9167	1.0000	1.0000	1.0000
madelon	0.5617	0.5767	0.5633	0.5833
ohsumed	0.9275	0.9300	0.9113	0.9438
apcj-etiology	0.9793	0.9826	0.9803	0.9879
dorothea	0.9200	0.9200	0.8857	0.9371
thrombin	0.9374	0.9429	0.9650	0.7300
news20	0.7755	-	-	0.7354
url1	0.9627	-	-	0.9765
kdd10	0.8780	-	-	0.8779
webspam	0.9532	-	-	0.9617
w/t/l	-	2/5/3	4/4/2	3/5/6

Table XII. Prediction accuracy (SVM)

Dataset	SAOLA	FCBF	SPSF-LAR	GDM
dexter	0.8500	0.5400	0.6400	0.9100
lung-cancer	0.9833	0.9800	0.9833	0.9833
hiva	0.9635	0.9635	0.9635	0.9661
breast-cancer	0.8750	0.8750	0.8854	0.4792
leukemia	0.9583	0.9167	0.9167	1.0000
madelon	0.6217	0.5933	0.7900	0.5833
ohsumed	0.9431	0.9431	0.9431	0.9438
apcj-etiology	0.9872	0.9872	0.9872	0.9879
dorothea	0.9286	0.9171	0.9029	0.9371
thrombin	0.9116	0.9116	0.9153	0.7300
news20	0.8721	-	-	0.7354
url1	0.9645	-	-	0.9765
kdd10	0.8727	-	-	0.8779
webspam	0.9123	-	-	0.9617
w/t/l	-	4/6/0	3/5/2	4/6/4

accordingly, the SAOLA and GDM do not have significant difference in prediction accuracy.

As for SVM, the null-hypothesis is rejected, and the average ranks for SAOLA and GDM are 1.25 and 1.75, respectively. Then we proceed with the Nemenyi test as a post-hoc test. With the Nemenyi test, the critical difference is up to 0.5238. Thus, with the critical difference and the average ranks calculated above, GDM is significantly better

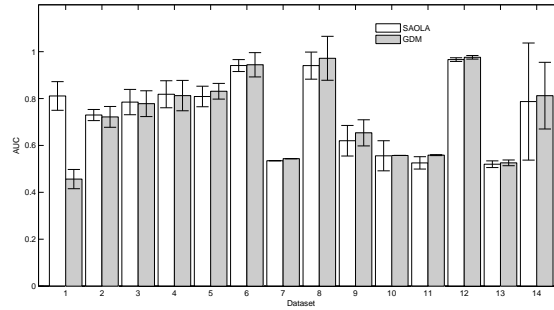


Fig. 10. The AUC of SAOLA and GDM (The labels of the x-axis from 1 to 14 denote the data sets: 1. breast-cancer; 2. dorothea; 3. thrombin; 4. news20; 5. dexter; 6. lung-cancer; 7. hiva; ; 8. leukemia; 9. madelon; 10. ohsumed; 11. apcj-etiology; 12. url1; 13. kdd10; 14. webspam)

than SAOLA. From the results above, we can see that the GDM algorithm is inferior to SAOLA on some data sets, such as *thrombin* and *news20*, since those data sets are very sparse. However, Fisher’s Z-test and information gain employed by SAOLA can deal with those spare data sets well.

In summary, our SAOLA algorithm is a scalable and accurate online approach. Without requiring a complete set of features on a training data set before feature selection starts, SAOLA is very competitive with the well-established and state-of-the-art FCBF, SPSF-LAR, and GDM methods.

5.3.2. AUC and Error Bar of SAOLA. In this section, we compare SAOLA with the three batch algorithms using the error bar and AUC metrics.

Table XIII. AUC of SAOLA, FCBF, and SPSF-LAR

Dataset	SAOLA	FCBF	SPSF-LAR
dexter	0.8088	0.7311	0.7611
lung-cancer	0.9407	0.9475	0.9639
hiva	0.5349	0.5347	0.5335
breast-cancer	0.8111	0.8120	0.8466
leukemia	0.9404	0.9737	0.9737
madelon	0.5972	0.5962	0.7537
ohsumed	0.5559	0.5741	0.5613
apcj-etiology	0.5256	0.5473	0.4987
dorothea	0.7297	0.6775	0.7218
thrombin	0.7849	0.7938	0.8382
average rank	1.8	1.95	2.15
w/t/l	-	3/4/3	2/3/5

Table XIII reports the average AUC of J48, KNN and SVM for each algorithm. Using the w/t/l counts, we can see that our SAOLA is very competitive with FCBF and SPSF-LAR. To further validate with the Friedman test and the null-hypothesis, the average ranks calculated for SAOLA, FCBF and SPSF-LAR are 1.8, 1.95, and 2.15, respectively. Accordingly, the AUC of SAOLA and that of FCBF and SPSF-LAR have no significant difference.

Figure 10 shows the average AUC of J48, KNN and SVM and its standard deviation for SAOLA and GDM. From Figure 10, GDM outperforms SAOLA on most data sets, but the AUC of SAOLA is highly comparable to that of GDM. Moreover, we can see that none of these algorithms, SAOLA, FCBF, SPSF-LAR, and GDM, can effectively deal with highly class-imbalanced data sets.

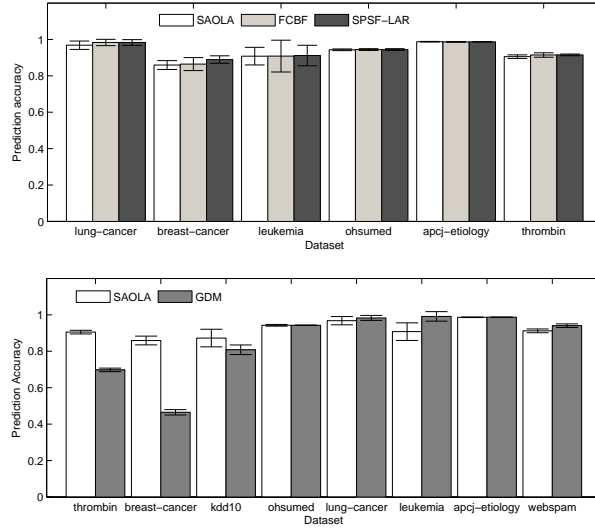


Fig. 11. The error bars of SAOLA, FCBF, SPSF-LAR, and GDM

We give the error bars of SAOLA, FCBF, SPSF-LAR, and GDM using SVM classifiers (on the *lung-cancer*, *breast-cancer*, *leukemia*, *ohsumed*, *apcj-etiology*, *thrombin*, *kdd10*, and *webspam* data sets) as shown in Figure 11. From Figure 11, SAOLA is very competitive with FCBF. Although GDM and SPSF-LAR achieve a higher prediction accuracy and a lower standard deviation than SAOLA, our SAOLA is highly comparable with those two batch methods.

5.4. Comparison with Markov Blanket Discovery Algorithms

In this section, we compare SAOLA and FCBF (discovery of approximate Markov blankets) with two state-of-the-art exact Markov blanket discovery algorithms, IAMB (Incremental Association Markov Blanket) [Tsamardinos and Aliferis 2003] and MMBB (Max-Min Markov Blanket) [Tsamardinos et al. 2006]. The IAMB algorithm finds Markov blankets conditioned on the selected feature set currently, while the MMBB algorithm discovers Markov blankets conditioned on all possible feature subsets of the selected feature set currently⁶. Under certain assumptions (sufficient number of data instances and reliably statistical tests), IAMB and MMBB can return the Markov blanket of a given target feature [Peña et al. 2007; Aliferis et al. 2010]. Using the four NIPS2003 feature selection challenge data sets, *arcene* (100 data instances, 10,000 features), *dexter* (300 data instances, 20,000 features), *dorothea* (800 data instances, 100,000 features), and *gisette* (6,000 data instance, 5,000 features), we empirically study SAOLA, FCBF, IAMB, and MMBB using SVM, KNN, and J48.

From Table XIV to Table XVI, we can see that using small sample-to-feature ratio data sets, such as *arcene*, *dexter*, *dorothea*, SAOLA and FCBF are competitive with IAMB and MMBB, and even better than IAMB and MMBB sometimes. The explanation is that the number of data instances required by IAMB to identify a Markov blanket is at least exponential in the size of the Markov blanket since IAMB considers a conditional independence test to be reliable when the number of data instances in D is at least five times the number of degrees of freedom in the test. MMBB mitigates

⁶We implement IAMB and MMBB using the package of Causal Explorer at http://www.dsl-lab.org/causal_explorer/.

Table XIV. Prediction and AUC of SAOLA, FCBF, IAMB, and MMB (SVM)

Dataset	Prediction Accuracy				AUC			
	SAOLA	FCBF	IAMB	MMMB	SAOLA	FCBF	IAMB	MMMB
arcene	0.6600	0.5600	0.6000	0.6700	0.6526	0.500	0.5917	0.6664
dorothea	0.9286	0.9171	0.9343	-	0.6455	0.5735	0.7799	-
dexter	0.8500	0.5400	0.7600	0.8467	0.8500	0.5400	0.7600	0.8467
gisette	0.8950	0.554	0.9340	0.9820	0.8950	0.5540	0.9340	0.9820

Table XV. Prediction and AUC of SAOLA, FCBF, IAMB, and MMB (KNN)

Dataset	Prediction Accuracy				AUC			
	SAOLA	FCBF	IAMB	MMMB	SAOLA	FCBF	IAMB	MMMB
arcene	0.6900	0.5900	0.5800	0.6200	0.6867	0.5804	0.5714	0.6096
dorothea	0.9200	0.9200	0.9143	-	0.7063	0.6932	0.6113	-
dexter	0.7600	0.7967	0.6600	0.8233	0.7600	0.7967	0.6600	0.8233
gisette	0.8600	0.8920	0.9300	0.9690	0.8600	0.8920	0.9300	0.9690

Table XVI. Prediction and AUC of SAOLA, FCBF, IAMB, and MMB (J48)

Dataset	Prediction Accuracy				AUC			
	SAOLA	FCBF	IAMB	MMMB	SAOLA	FCBF	IAMB	MMMB
arcene	0.5600	0.5500	0.5800	0.6000	0.5463	0.5252	0.5714	0.5966
dorothea	0.9343	0.9314	0.9371	-	0.7536	0.7258	0.7683	-
dexter	0.8133	0.8567	0.7833	0.8767	0.8133	0.8567	0.7833	0.8767
gisette	0.8960	0.9130	0.9260	0.9420	0.8960	0.9130	0.9260	0.9420

this problem to some extent. But the performance of IAMB and MMB may be inferior to SAOLA and FCBF as the sample-to-feature ratio becomes very small. Furthermore, as the size of a Markov blanket becomes large, MMB is very slow due to expensive computation costs, such as on the *dorothea* data set (the running time exceeds three days). But on the *gisette* data set, due to the large number of data instances, IAMB and MMB are significantly better than SAOLA and FCBF, but MMB and IAMB take much more running time, especially MMB. Those empirical results conclude that the performance of SAOLA is very close to that of IAMB and MMB on small sample-to-feature ratio data sets, but SAOLA is much more scalable than IAMB and MMB on data sets with both many data instances and extremely high dimensionality.

Table XVII. Number of selected features and running time of SAOLA, FCBF, IAMB, and MMB

Dataset	Number of selected features				Running time			
	SAOLA	FCBF	IAMB	MMMB	SAOLA	FCBF	IAMB	MMMB
arcene	22	31	3	5	3	3	5	12
dorothea	63	96	6	-	58	78	479	-
dexter	21	55	4	11	3	5	38	74
gisette	22	37	9	308	10	8	131	13,483

5.5. Analysis of the Effect of Parameters on SAOLA

5.5.1. Analysis of Correlation Bounds. In Section 3.3, we set the derived the correlation bound of $I(F_i; Y)$ to $\min(I(F_i; C), I(Y; C))$. In this section, we investigate SAOLA using the opposite bound, i.e., $\max(I(F_i; C), I(Y; C))$, which we term the SAOLA-max algorithm. In the experiments, SAOLA-max uses the same parameters as SAOLA.

Table XVIII shows the prediction accuracies of SAOLA and SAOLA-max. With the summary of the w/t/l counts in Table XVIII, we can see that SAOLA is very competitive with SAOLA-max in prediction accuracy. With the Friedman test at 95% significance level, As for SVM, SAOLA gets the higher average rank than SAOLA-max. For KNN, the null-hypothesis cannot be rejected. The average ranks calculated from the Friedman test for SAOLA and SAOLA-max are 1.46 and 1.54, respectively. With respect to

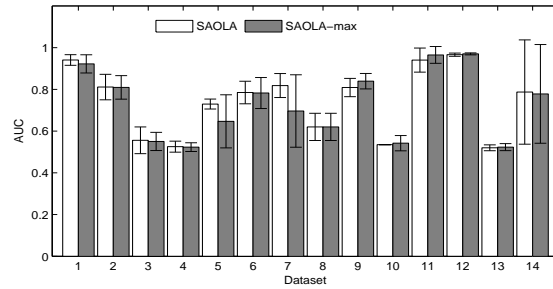


Fig. 12. The AUC of SAOLA and SAOLA-max (The labels of the x-axis from 1 to 14 denote the data sets: 1. lung-cancer; 2. breast-cancer; 3. ohsumed; 4. apcj-etiology; 5. dorothea; 6. thrombin; 7. news20; 8. madelon; 9. dexter; 10. hiva; 11. leukemia; 12. url1; 13. kdd10; 14. webspam)

J48, the average ranks for SAOLA and SAOLA-max are 1.43 and 1.57, respectively. The Friedman test testifies that SAOLA and SAOLA-max have no significant difference in prediction accuracy, although SAOLA-max gets the higher average ranks using the J48 and KNN classifiers.

Table XVIII. Prediction accuracy

Dataset	KNN		J48		SVM	
	SAOLA	SAOLA-max	SAOLA	SAOLA-max	SAOLA	SAOLA-max
dexter	0.7600	0.8000	0.8133	0.8300	0.8500	0.8800
lung-cancer	0.9833	0.9500	0.9500	0.9500	0.9833	1.0000
hiva	0.9635	0.9505	0.9661	0.9557	0.9635	0.9635
breast-cancer	0.8646	0.8854	0.7917	0.8125	0.8750	0.8645
leukemia	0.9167	1.0000	0.9583	0.9583	0.9583	0.8750
madelon	0.5617	0.5617	0.6083	0.6083	0.6217	0.6217
ohsumed	0.9275	0.9256	0.9437	0.9437	0.9431	0.9431
apcj-etiology	0.9793	0.9807	0.9872	0.9870	0.9872	0.9872
dorothea	0.9200	0.9171	0.9343	0.9257	0.9286	0.9029
thrombin	0.9374	0.9484	0.9613	0.9503	0.9116	0.9153
news20	0.7755	0.7592	0.8276	0.8295	0.8721	0.4993
url1	0.9627	0.9732	0.9744	0.9761	0.9645	0.9614
kdd2010	0.8780	0.8766	0.8723	0.8751	0.8727	0.8727
webspam	0.9532	0.9546	0.9611	0.9635	0.9123	0.8798
Ave rank	1.46	1.54	1.43	1.57	1.61	1.39
w/t/l	-	4/5/5	-	2/10/2	-	5/7/2

Moreover, Figure 12 shows the average AUC of J48, KNN and SVM and its standard deviation for SAOLA and SAOLA-max. We can see that SAOLA and SAOLA-max are very competitive with each other on all 14 data sets.

However, on the running time, Table XIX shows that SAOLA is much more efficient than SAOLA-max on all data sets, especially on those of extremely high dimensionality. In Table XIX, we can also see that SAOLA selects fewer features than SAOLA-max. The explanation is that SAOLA-max uses a bigger relevance threshold ($\delta_2 = \max(I(X; C), I(Y; C))$) for removing redundant features than SAOLA ($\delta_2 = \min(I(X; C), I(Y; C))$). Clearly, the larger the relevance threshold δ_2 , more features are added to the current feature set (see Steps 11 and 16 of Algorithm 1).

Compared to SAOLA-max, we can conclude that it is accurate and scalable to use the correlation bound, $\delta_2 = \min(I(X; C), I(Y; C))$ in the SAOLA algorithm, for pairwise comparisons to filter out redundant features.

Finally, we evaluate the stabilities of SAOLA and SAOLA-max using the stability measure proposed by Yu et al. [Yu et al. 2008]. Each data set was randomly partitioned

Table XIX. Running time and Number of selected features

Dataset	Running time (seconds)		Number of selected features	
	SAOLA	SAOLA-max	SAOLA	SAOLA-max
dexter	3	3	21	39
lung-cancer	6	62	35	260
hiva	1	3	12	58
breast-cancer	5	40	46	93
leukemia	2	4	17	70
madelon	0.1	0.1	3	3
ohsumed	6	8	65	89
apcj-etiology	22	38	75	105
dorothea	58	327	63	516
thrombin	63	497	20	498
news20	944	2,100	212	449
url1	200	526	64	346
kdd2010	1,056	2,651	180	193
webspam	1,456	11,606	51	165

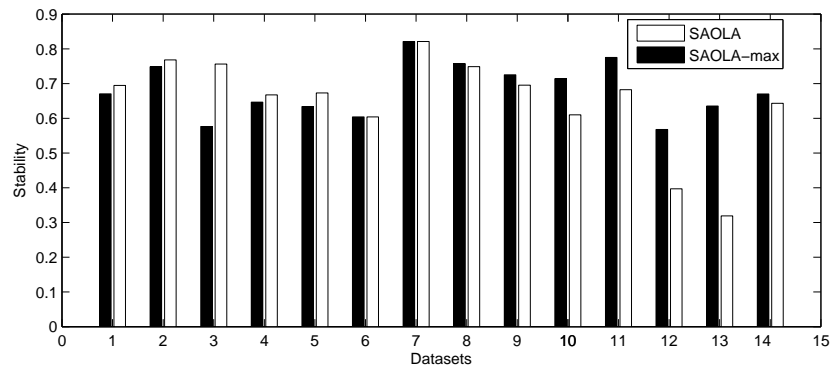


Fig. 13. The stabilities of SAOLA and SAOLA-max (The labels of the x-axis from 1 to 14 denote the data sets: 1. dexter; 2. lung-cancer; 3. ohsumed; 4. apcj-etiology; 5. news20; 6. breast-cancer; 7. madelon; 8. leukemia; 9. kdd10; 10. webspam; 11. hiva; 12. dorothea; 13. thrombin; 14. url)

into five folds, each fold containing 1/5 of all the samples. SAOLA and SAOLA-max were repeatedly applied to four out of the five folds. This process was repeated 30 times to generate different subsamples for each data set. Then the average stabilities over 30 subsamples on each data set are as the results of SAOLA and SAOLA-max. Figure 13 shows the stabilities of SAOLA and SAOLA-max. We can conclude that SAOLA is very competitive with SAOLA-max on the measure of stability, although SAOLA-max can select large subsets of features.

5.5.2. The Effect of Input Order of Features. Since the dimensions are presented in a sequential scan, does the input order of the features have an impact on the quality of the selected feature set? To evaluate the effect on the SAOLA algorithm, we generate 30 trials in which each trial represents a random ordering of the features in the input feature set. We apply the SAOLA algorithm to each randomized trial and report the average prediction accuracies and standard deviations ($\text{accuracy} \pm \text{deviation}$) in Table XX (in the following sections, we only give the prediction accuracy of SAOLA using KNN and J48, since the prediction accuracy of SAOLA using SVM is very similar to that of SAOLA using KNN.).

On the last eight very high-dimensional data sets, the results in Table XX confirm that varying the order of the incoming features does not affect much the final outcomes.

Table XX. Average prediction accuracy and standard deviation of SAOLA

Dataset	KNN (accuracy±deviation)	J48 (accuracy±deviation)
ohsumed	0.9389±0.0039	0.9444±0.0002
apcj-etiology	0.9824±0.0008	0.9871±0.0002
dorothea	0.9193±0.0095	0.9345±0.0044
thrombin	0.9498±0.0072	0.9501±0.0021
kdd10	0.8759±0.0025	0.8682±0.0066
news20	0.7694±0.0052	0.8194±0.0041
url1	0.9557±0.0107	0.9729±0.0023
webspam	0.9502±0.0035	0.9618±0.0035

Our explanation is that with various feature orders, Steps 11 and 16 of Algorithm 1 can select the feature with the highest correlation with the class attribute among a set of correlated features and remove the corresponding correlated features of this feature.

The only difference is that in some feature orders, the final feature subset may include some weakly relevant features. For example, assuming at time t , F_i arrives and has only one feature Y that satisfies Eq.(18) in the input features, and Y arrived before F_i and has stayed in the currently selected feature set $S_{t_{i-1}}^*$. Then F_i can be removed at time t given Y . But if F_i arrives before Y , and Y is removed before F_i 's arrival, F_i cannot be removed later and may be kept in the final feature set. This also explains why there is a little fluctuation of standard deviations in Table XX.

Table XXI. Prediction Accuracy

Dataset	KNN		J48	
	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$
madelon	0.5617	0.5717	0.6083	0.5416
ohsumed	0.9275	0.9394	0.9437	0.9437
apcj-etiology	0.9793	0.9838	0.9872	0.9873
news20	0.7755	0.7749	0.8276	0.8276
url1	0.9627	0.9642	0.9744	0.9744
kdd2010	0.8780	0.8678	0.8723	0.8723
webspam	0.9532	0.9493	0.9611	0.9611

5.5.3. The Effect of δ and α . The SAOLA algorithm has two versions: SAOLA with information gain for discrete data and SAOLA with Fisher's Z-test for continuous data. For both versions, SAOLA needs to set a relevance threshold (δ in Algorithm 1) in advance to determine whether two features are relevant. For discrete data, we set 11 different relevance thresholds for SAOLA and tuned δ using cross validation on the *dorothea* and *thrombin* data sets. From Figure 14, we can see that in the term of prediction accuracy, the relevance thresholds do not have a significant impact on the SAOLA algorithm.

For Fisher's Z-test, the relevance threshold is the significance level, α , and is always set to 0.01 or 0.05. Table XXI shows the results of SAOLA under the different significance levels. It is clear that a significant level does not impose a significant impact on the SAOLA algorithm either.

5.6. Comparison of Group-SAOLA with OGFS and Sparse Group Lasso

In this section, we compare our group-SAOLA algorithm with the state-of-the-art online group feature selection methods, OGFS [Wang et al. 2013] and a well-established batch group feature selection method, Sparse Group Lasso [Friedman et al. 2010]. As for the first ten high-dimensional data sets in Table II, from *dexter* to *thrombin*, we randomly separated each data set into 100 feature groups without overlapping, while for the last four data sets with extremely high dimensionality, each data set was randomly separated into 10,000 feature groups without overlapping. For parameter settings of

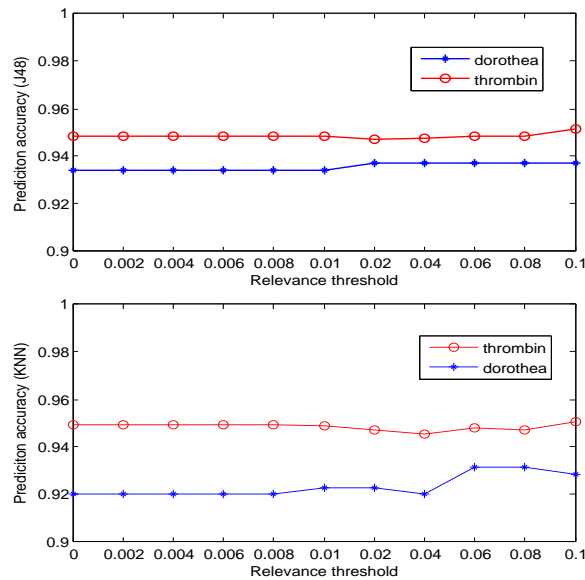


Fig. 14. Prediction accuracies on varied relevance thresholds (the top figure with J48 while the bottom figure with KNN)

the Sparse Group Lasso algorithm, $\lambda_1 \in [0.01, 0.1]$ and $\lambda_2 \in [0.01, 0.1]$. For the group-SAOLA algorithm, the parameters, δ for discrete data and α for continuous data, are the same as the SAOLA algorithm. In the following comparisons, since the prediction accuracy of SAOLA using SVM has no significant difference than that of SAOLA using KNN and J48, we only report the prediction accuracy of SAOLA using KNN and J48.

In this section we compare group-SAOLA with OGFS and Sparse Group Lasso in terms of prediction accuracy, sizes of selected feature subsets, number of selected groups, and running time on the 14 high-dimensional data sets in Table II. We repeated this process 10 times to generate different sets of feature groups of each data set. The results as shown in Table XXII, Table XXIII, Figures 15 to 18 are the average ones on those 10 sets of feature groups.

Table XXII summarizes the prediction accuracies of Group-SAOLA against OGFS and Sparse Group Lasso using the KNN and J48 classifiers. The highest prediction accuracy is highlighted in bold face. Table XXIII illustrates the running time, sizes of selected feature subsets, and numbers of selected groups of Group-SAOLA against OGFS and Sparse Group Lasso. In Tables XXII and XXIII, SGLasso is the abbreviation for Sparse Group Lasso.

5.6.1. Comparison of Group-SAOLA with OGFS. In this section we compare OGFS with group-SAOLA on the 14 high-dimensional data sets in Table II, and the results are as shown in Tables XXII to XXIII, Figures 15 to 18. In Tables XXII, from the the win/tie/lose counts in the last rows of the table, we observe that Group-SAOLA never loses against OGFS on all of the 14 high-dimensional data sets.

To evaluate whether the prediction accuracy of group-SAOLA and that of OGFS have no significant difference, using the Friedman test, for the J48 classifier, the null-hypothesis is rejected, and the average ranks for group-SAOLA and OGFS are 1.8929 and 1.1071, respectively. Then we proceed with the Nemenyi test as a post-hoc test, and the critical difference is up to 0.6885. The difference between 1.8929 and 1.1071

Table XXII. Prediction Accuracy

Dataset	J48			KNN		
	group-SAOLA	OGFS	SGLasso	group-SAOLA	OGFS	SGLasso
dexter	0.8427	0.5556	0.8800	0.7947	0.5487	0.7067
lung-cancer	0.9500	0.9017	0.9833	0.9584	0.9167	0.9333
hiva	0.9661	0.9602	0.9609	0.9630	0.9471	0.9479
breast-cancer	0.6656	0.6000	0.6667	0.6531	0.6385	0.6667
leukemia	0.9583	0.7292	0.9583	0.9833	0.7792	1.0000
madelon	0.6117	0.5153	0.6533	0.5317	0.4922	0.5967
ohsumed	0.9439	0.9430	0.9431	0.9052	0.9142	0.9431
apcj-etiology	0.9872	0.9872	0.9872	0.9788	0.9790	0.9818
dorothea	0.9365	0.9029	0.9314	0.9183	0.8691	0.9143
thrombin	0.9591	0.9396	0.9558	0.9376	0.9420	0.9632
news20	0.8188	0.5303	-	0.7501	0.5058	-
url	0.9715	0.6333	-	0.9553	0.6089	-
kdd10	0.8714	0.8787	-	0.8764	0.8758	-
webspam	0.9341	0.7620	-	0.9376	0.9376	-
w/t/l	-	10/4/0	0/7/3	-	9/5/0	3/2/5

Table XXIII. Running time and Number of selected features/groups

Dataset	Running time (seconds)			Number of selected features/groups		
	group-SAOLA	OGFS	SGLasso	group-SAOLA	OGFS	SGLasso
dexter	2	4	69	21/19	72/49	56/40
lung-cancer	9	2	1,752	23/19	91/61	384/41
hiva	2	1	734	5/5	47/38	55/19
breast-cancer	8	3	275	8/7	111/61	2,775/30
leukemia	3	1	18	16/14	66/47	61/28
madelon	0.1	0.1	27	2/2	15/13	5/2
ohsumed	2	6	64	17/16	61/43	385/28
apcj-etiology	34	34	210	47/33	44/44	44/12
dorothea	23	22	112	41/27	126/67	135/6
thrombin	39	1,015	1,015	7/5	691/84	691/70
news20	2,154	1,054	-	140/140	192/192	-
url	306	3,598	-	29/29	73/73	-
kdd10	395	39,213	-	52/52	133/132	-
webspam	3,013	20,718	-	17/17	401/395	-

is bigger than this critical difference, then group-SAOLA is significantly better than OGFS in prediction accuracy.

For the KNN classifier, the null-hypothesis cannot be rejected. The average ranks for group-SAOLA and OGFS are 1.75 and 1.25, respectively. Accordingly, for KNN, group-SAOLA and OGFS have no significant difference in prediction accuracy.

Figure 15 gives the error bars (the left figure) and AUC (the right figure) of group-SAOLA and OGFS using the KNN classifier. We can see that group-SAOLA clearly outperforms OGFS using the AUC and error bar metrics.

Furthermore, Table XXIII illustrates that group-SAOLA is faster than OGFS on most data sets. When the number of features increases to millions and the number of feature groups becomes large, OGFS becomes very costly, but our group-SAOLA is still scalable and efficient. The explanation is that the time complexity of group-SAOLA is determined by the number of features within the currently selected feature groups, and the strategy of online redundancy detection within the currently selected feature groups makes group-SAOLA very scalable. Meanwhile, from Table XXIII, we observe that group-SAOLA selects fewer features than OGFS. From the selected numbers of groups and selected numbers of features, we can see that group-SAOLA not only selects the smaller number of feature groups, but also achieves more sparsity of within groups than OGFS.

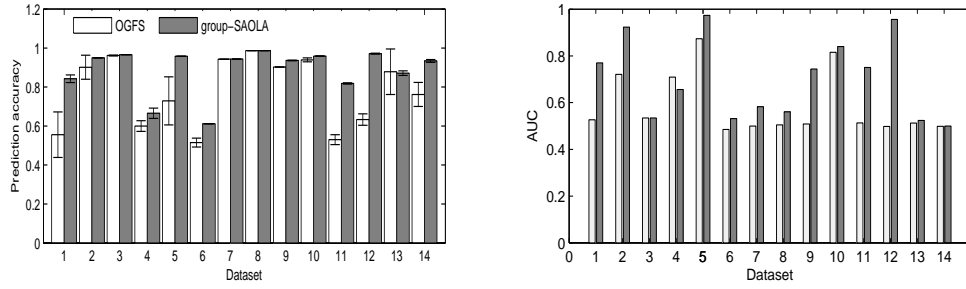


Fig. 15. The error bars and AUC of group-SAOLA and OGFS (The labels of the x-axis from 1 to 14 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin; 11. news20; 12. url1; 13. kdd10; 14. webspam)

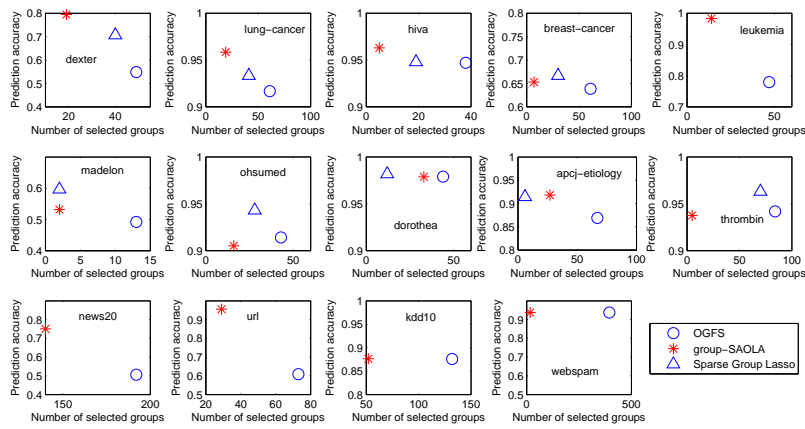


Fig. 16. Accuracies and numbers of selected groups of three algorithms (KNN)

Figure 16 gives the results of the numbers of selected groups and the corresponding prediction accuracy for group-SAOLA, and OGFS using the KNN classifier on the 14 data sets in Table II. The best possible mark for each graph is at the upper left corner, which selects the fewest groups with the highest prediction accuracy. We can see that group-SAOLA selects fewer groups while gets higher prediction accuracies than OGFS on all the 14 data sets in Table II, except for the *ohsumed* data set. However, on the *ohsumed* data set, on the prediction accuracy, group-SAOLA is very competitive with OGFS.

Figure 17 gives the results of the numbers of selected groups and the corresponding prediction accuracy for Group-SAOLA and OGFS using the J48 classifier. We can see that Group-SAOLA selects fewer groups while gets higher prediction accuracies than OGFS on all the fourteen data sets.

5.6.2. Comparison of Group-SAOLA with Sparse Group Lasso. Since Sparse Group Lasso⁷ can only deal with the first ten high-dimensional data sets in Table II due to high computational costs, in this section we compare it with group-SAOLA on those first ten high-dimensional data sets.

⁷The codes are available at <http://yelab.net/software/SLEP/>

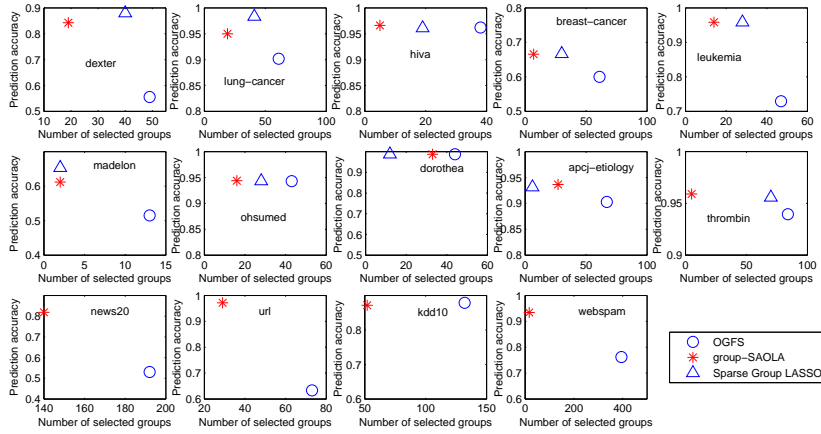


Fig. 17. Accuracies and numbers of selected groups of three algorithms (J48)

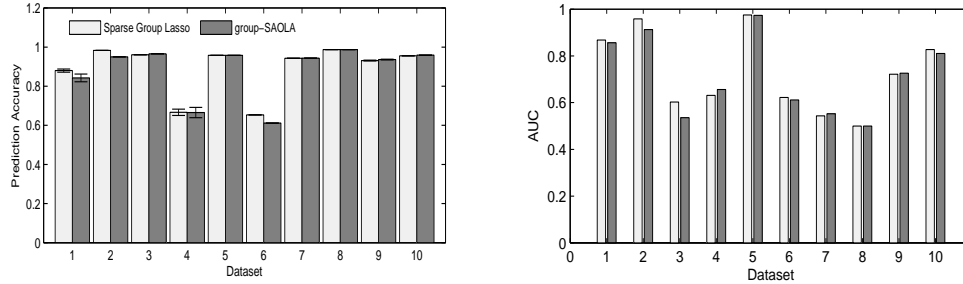


Fig. 18. The error bars and AUC of group-SAOLA and Sparse Group Lasso (The labels of the x-axis from 1 to 10 denote the data sets: 1. dexter; 2. lung-cancer; 3. hiva; 4. breast-cancer; 5. leukemia; 6. madelon; 7. ohsumed; 8. apcj-etiology; 9. dorothea; 10. thrombin)

With Table XXII, using the Friedman test, for the J48 classifier, the null-hypothesis cannot be rejected, and the average ranks for group-SAOLA and Sparse Group Lasso are 1.5 and 1.5, respectively. For the KNN classifier, the null-hypothesis is accepted, and the average ranks for group-SAOLA and Sparse Group Lasso are 1.6 and 1.4, respectively. Accordingly, for the J48 and KNN classifiers, group-SAOLA and Sparse Group Lasso have no significant difference in prediction accuracy. Furthermore, Table XXIII shows that group-SAOLA is much faster than Sparse Group Lasso, and group-SAOLA selects fewer features than Sparse Group Lasso.

Figure 18 gives the AUC (the left figure) and error bars (the right figure) of group-SAOLA and Sparse Group Lasso using the KNN classifier. Figure 18 illustrates that group-SAOLA is very competitive with Sparse Group Lasso using the AUC and error bar metrics.

Figures 16 and 17 give the results of the numbers of selected groups and the corresponding prediction accuracy for group-SAOLA and Sparse Group Lasso using the KNN and J48 classifiers, respectively. The best possible mark for each graph is at the upper left corner. We can see that group-SAOLA prefers to select few groups, in comparison to Sparse Group Lasso. Meanwhile, group-SAOLA is very competitive with Sparse Group Lasso in terms of prediction accuracy.

In summary, the group-SAOLA algorithm is a scalable and accurate online group feature selection approach. This validates that without requiring a complete set of feature groups on a training data set before feature selection starts, group-SAOLA is very competitive comparing to the well-established the Sparse Group Lasso algorithm.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the SAOLA algorithm, a scalable and accurate online approach to tackle feature selection with extremely high dimensionality. We conducted a theoretical analysis and derived a lower bound of correlations between features for pairwise comparisons, and then proposed a set of online pairwise comparisons to maintain a parsimonious model over time. To deal with the group structure information in features, we extended the SAOLA algorithm, and then proposed a novel group-SAOLA algorithm to deal with features that arrive by groups. The group-SAOLA algorithm can online maintain a set of feature groups that is sparse between groups and within each group simultaneously.

Using a series of benchmark data sets, we compared the SAOLA and group-SAOLA algorithms with state-of-the-art online feature selection methods and well-established batch feature selection algorithms. The empirical study demonstrated that the SAOLA and group-SAOLA algorithms are both scalable on data sets of extremely high dimensionality, have superior performance over state-of-the-art online feature selection methods, and are very competitive with state-of-the-art batch feature selection methods in prediction accuracy, while much faster in running time.

In this work, we have used online pairwise comparisons to calculate the correlations between features without further exploring positive feature interactions between features. Moreover, from the AUC results reported in the work, we can see that SAOLA and its rivals, including the three online algorithms and three batch methods, cannot effectively deal with class-imbalanced data. Thus, we will further explore the following directions in online feature selection: efficient and effective methods to discover positive feature interactions between features, and accurate and scalable online algorithms to handle class-imbalanced data.

Meanwhile, our empirical studies have validated that SAOLA (using pairwise comparisons) is competitive with IAMB and MMB (using multiple comparisons). To conduct thoroughly theoretical analysis and empirical studies on why pairwise feature correlations (instead of conditioning on all possible feature subsets) may be sufficient in practice deserve further exploration in our future work.

Acknowledgments

We thank the handling editor and the anonymous reviewers for their invaluable and constructive comments, which helped to improve the paper substantially. This work is partly supported by the National 973 Program of China (under grant 2013CB329604), a PIMS Post-Doctoral Fellowship Award of the Pacific Institute for the Mathematical Sciences, Canada, the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China (under grant IRT13059), the National Natural Science Foundation of China (under grants 61229301 and 61305064), an NSERC Discovery grant and a BCIC NRAS Team Project.

All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 2010. Local causal and markov blanket induction for causal discovery and feature selection for

- classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research* 11 (2010), 171–234.
- Gavin Brown, Adam Pockock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research* 13 (2012), 27–66.
- Zhiguang Chen, Yutong Lu, Nong Xiao, and Fang Liu. 2014. A hybrid memory built by SSD and DRAM to support in-memory Big Data analytics. *Knowledge and Information Systems* 41, 2 (2014), 335–354.
- Manoranjan Dash and Huan Liu. 2003. Consistency-based search in feature selection. *Artificial intelligence* 151, 1 (2003), 155–176.
- Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7 (2006), 1–30.
- George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research* 3 (2003), 1289–1305.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2010. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736* (2010).
- Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (2003), 1157–1182.
- Steven CH Hoi, Jialei Wang, Peilin Zhao, and Rong Jin. 2012. Online feature selection for mining big data. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. ACM, 93–100.
- Aleks Jakulin and Ivan Bratko. 2003. Analyzing attribute dependencies. In *PKDD 2003*. Springer-Verlag, 229–240.
- Kashif Javed, Mehreen Saeed, and Haroon A Babri. 2014. The correctness problem: evaluating the ordering of binary features in rankings. *Knowledge and information systems* 39, 3 (2014), 543–563.
- Alexandros Kalousis, Julien Prados, and Melanie Hilario. 2007. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems* 12, 1 (2007), 95–116.
- Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1 (1997), 273–324.
- Daphne Koller and Mehran Sahami. 1995. Toward optimal feature selection. In *ICML-1995*. 284–292.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* (1951), 79–86.
- Bo Liu, Yanshan Xiao, S Yu Philip, Zhifeng Hao, and Longbing Cao. 2014. An efficient orientation distance-based discriminative feature extraction method for multi-classification. *Knowledge and information systems* 39, 2 (2014), 409–433.
- Huan Liu and Lei Yu. 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering* 17, 4 (2005), 491–502.
- Jose M Peña. 2008. Learning gaussian graphical models of gene networks with false discovery rate control. In *EvoBIO-2008*. 165–176.
- Jose M Peña, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. 2007. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning* 45, 2 (2007), 211–232.
- Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 8 (2005), 1226–1238.
- Simon Perkins and James Theiler. 2003. Online feature selection using grafting. In *ICML*. 592–599.
- William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. 1996. *Numerical recipes in C*. Vol. 2. Citeseer.
- Jonathon Shlens. 2014. Notes on Kullback-Leibler Divergence and Likelihood. *arXiv preprint arXiv:1404.2000* (2014).
- Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13 (2012), 1393–1434.
- Mingkui Tan, Ivor W Tsang, and Li Wang. 2014. Towards Ultrahigh Dimensional Feature Selection for Big Data. *Journal of Machine Learning Research* 15 (2014), 1371–1429.
- Mingkui Tan, Li Wang, and Ivor W Tsang. 2010. Learning sparse svm for feature selection on very high dimensional datasets. In *ICML-2010*. 1047–1054.
- Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

- Ioannnis Tsamardinos and Constantin F Aliferis. 2003. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the ninth international workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann Publishers: Key West, FL, USA.
- Ioannnis Tsamardinos, Laura E Brown, and Constantin F Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning* 65, 1 (2006), 31–78.
- De Wang, Danesh Irani, and Calton Pu. 2012. Evolutionary Study of Web Spam: Webb Spam Corpus 2011 versus Webb Spam Corpus 2006. In *CollaborateCom-2012*. 40–49.
- Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin. 2013. Online Feature Selection and Its Applications. *IEEE Transactions on Knowledge and Data Engineering* (2013), 1–14.
- Jason Weston, Sayan Mukherjee, Olivier Chapelle, Massimiliano Pontil, Tomaso Poggio, and Vladimir Vapnik. 2000. Feature selection for SVMs. In *NIPS*, Vol. 12. 668–674.
- Adam Woznica, Phong Nguyen, and Alexandros Kalousis. 2012. Model mining for robust feature selection. In *ACM SIGKDD-2012*. ACM, 913–921.
- Xindong Wu, Kui Yu, Wei Ding, Hao Wang, and Xingquan Zhu. 2013. Online feature selection with streaming features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013), 1178–1192.
- Xindong Wu, Kui Yu, Hao Wang, and Wei Ding. 2010. Online streaming feature selection. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 1159–1166.
- Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. 2014. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on* 26, 1 (2014), 97–107.
- Jin Xiao, Yi Xiao, Annqiang Huang, Dunhu Liu, and Shouyang Wang. 2015. Feature-selection-based Dynamic Transfer Ensemble Model for Customer Churn Prediction. *Knowledge and Information Systems* 43, 1 (2015), 29–51.
- Kui Yu, Wei Ding, Dan A Simovici, Hao Wang, Jian Pei, and Xindong Wu. 2015a. Classification with Streaming Features: An Emerging-Pattern Mining Approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 9, 4 (2015), 30.
- Kui Yu, Dawei Wang, Wei Ding, Jian Pei, David L Small, Shafiqul Islam, and Xindong Wu. 2015b. Tornado Forecasting with Multiple Markov Boundaries. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2237–2246.
- Lei Yu, Chris Ding, and Steven Loscalzo. 2008. Stable feature selection via dense feature groups. In *ACM SIGKDD-2008*. ACM, 803–811.
- Lei Yu and Huan Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research* 5 (2004), 1205–1224.
- Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67.
- Yiteng Zhai, Y Ong, and I Tsang. 2014. The Emerging “Big Dimensionality”. *Computational Intelligence Magazine, IEEE* 9, 3 (2014), 14–26.
- Yiteng Zhai, Mingkui Tan, Ivor Tsang, and Yew Soon Ong. 2012. Discovering Support and Affiliated Features from Very High Dimensions. In *ICML-2012*. 1455–1462.
- Xiangrong Zhang, Yudi He, Licheng Jiao, Ruochen Liu, Ji Feng, and Sisi Zhou. 2015. Scaling Cut Criterion-based Discriminant Analysis for Supervised Dimension Reduction. *Knowledge and information systems* 43, 3 (2015), 633–655.
- Zheng Zhao and Huan Liu. 2007. Searching for Interacting Features.. In *IJCAI*, Vol. 7. 1156–1161.
- Zheng Zhao, Lei Wang, Huan Liu, and Jieping Ye. 2013. On similarity preserving feature selection. *IEEE Transactions on Knowledge and Data Engineering* 25 (2013), 619–632.
- Jing Zhou, Dean P Foster, Robert A Stine, and Lyle H Ungar. 2006. Streamwise feature selection. *Journal of Machine Learning Research* 7 (2006), 1861–1885.
- Tianyi Zhou, Dacheng Tao, and Xindong Wu. 2011. Manifold elastic net: a unified framework for sparse dimension reduction. *Data Mining and Knowledge Discovery* 22, 3 (2011), 340–371.