Lex-SVM: exploring the potential of exon expression profiling for disease classification

Xiongying Yuan
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
Graduate School of Chinese Academy of Sciences, Beijing 100039, China

Yi Zhao, Changning Liu and Dongbo Bu*
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
**E-mail: dbu@ict.ac.cn*

Exon expression profiling technologies, including exon arrays and RNA-Seq, measure the abundance of every exon in a gene. Compared with gene expression profiling technologies like 3' array, exon expression profiling technologies could detect alterations in both transcription and alternative splicing, therefore are expected to be more sensitive in diagnosis. However, exon expression profiling also brings higher dimension, more redundancy, and significant correlation among features. Ignoring the correlation structure among exons of a gene, popular classification method like L1-SVM selects exons individually from each gene and thus is vulnerable to noise. To overcome this limitation, we present in this paper a new variant of SVM named Lex-SVM to incorporate correlation structure among exons and known splicing patterns to promote classification performance. Specifically, we construct a new norm, ex-norm, including our prior knowledge on exon correlation structure to regularize the coefficients of a linear SVM. Lex-SVM can be solved efficiently using standard linear programming techniques. The advantage of Lex-SVM is that it can select features group-wisely, force features in a subgroup to take equal weights and exclude the features that contradict the majority in the subgroup. Experimental results suggest that on exon expression profile Lex-SVM is more accurate than existing methods. Lex-SVM also generates a more compact model and selects genes more consistently in cross-validation. Unlike L1-SVM selecting only one exon in a gene, Lex-SVM assigns equal weights to as many exons in a gene as possible, lending itself easier for further interpretation.

1. INTRODUCTION

Variable transcripts from a single gene are produced combinatorially through the selection of exons in alternative splicing (AS) [4]. AS is one of the most important sources of protein diversity in vertebrates. Recent bioinformatics analysis suggests that almost 70% of human genes are alternatively spliced [13]. AS is not only involved in normal development, but is also associated with human diseases including cancer [16, 3]. In order to detect alternative splicing, several platforms have been developed, e.g., Affymetrix exon arrays, and more recently, RNA-Seq. By measuring the abundance of exons in a variety of tissues, these platforms could test those diseases caused by aberration in transcription and alternative splicing, therefore are expected to be more sensitive in diagnosis than gene expression profiling technologies like 3' array.

On traditional 3' array data, SVM and its variants have demonstrated superior performance in classification and biomarker selection [7, 30]. Suppose we have n training samples (x_i, y_i) indexed by i. $x_i \in R^m$ contains all probeset measurements in array i, and $y_i \in \{1, -1\}$ defines the class of x_i. SVM searches for a hyperplane $f(X) = w^T x + w_0$ that maximizes the margin between the training samples in class 1 and class -1. Formally, the standard SVM can be described as follows:

$$\begin{aligned} & \max_{w,w_0} \frac{1}{\|w\|_2} \\ \text{s.t.} \quad & y_i(x_i^T w + w_0) \geq 1 - \xi_i \\ & \sum_{i=1}^{n} \xi_i \leq c, \qquad \xi_i \geq 0 \end{aligned}$$

, where ξ_i are slack variables and c is a tuning parameter. This standard SVM has an equivalent *hinge loss + penalty* formulation:

$$\min_{w,w_0} \sum_{i=1}^{n} [1 - y_i(x_i^T w + w_0)]_+ + \lambda \|w\|_2^2 \qquad (1)$$

, where the subscript '+' denotes the positive part, i.e. $z_+ = \max\{z, 0\}$. The 2-norm penalty helps con-

*Corresponding author.

trol the model complexity to prevent over-fitting on training data.

The difficulty in microarray data analysis is characterized by its *'high dimension but small sample size'*. Therefore, robust and accurate feature selection methods are required to identify the features differentially expressed across different samples, e.g., between cancerous and normal cells. The benefits of feature selection are two-fold: to provide a better understanding of the underlying biological system and to improve the classification performance. However, the standard SVM utilizes all the variables without discrimination, leading to its inability to conduct feature selection.

To accomplish the goal of automatic feature selection, L1-SVM [30] was proposed by replacing the 2-norm penalty with the lasso penalty [22]. It can be formulated as

$$\min_{w,w_0} \sum_{i=1}^{n} [1 - y_i(x_i^T w + w_0)]_+ + \lambda \|w\|_1 \qquad (2)$$

Following this framework, a variety of penalty functions were proposed for specific applications: To account for group-wise feature selection, Yuan and Lin [27] proposed the so-called *group lasso* in linear regression, and Zou and Yuan [31] later proposed *maximum norm* in F-∞ SVM for higher computational efficiency. In order to capture the correlation between successive variables, *variable fusion* penalty function was proposed to force successive variables of the classifier to have similar weights [14, 23]. This technique has been successfully used in arrayCGH data analysis [19].

Compared with gene expression profile, exon expression profile raises several new challenges to the above supervised classification techniques: i) the number of features has increased dozens of times from gene to exon expression profile. This exacerbates the *'high dimension but small sample size'* difficulty in microarray analysis. ii) the expression of a gene's exons are highly correlated unless alternative splicing exists.

If applied on exon expression profile, the above methods have to either treat all the exons of a gene as though they belonged to the same single transcript, effectively averaging, or treat all the exons as independent entities with no correlation structure. The first approach will not be able to diagnose diseases for which the expression level of some alternative exon(s) is a key indicator, but the average expression over all exons in the gene is not. The second approach fails to pool data from exons that are always expressed together, decreasing the sample size for estimation of transcript abundance and hence potentially increasing sampling error. To deal with this problem, we developed a new variant of SVM – Lex-SVM to incorporate known splicing patterns into classification.

2. METHODS

In this section, we first elaborate the exon expression correlation structure contained in sequence data, and then incorporate this correlation structure to generate a new variant of SVM – Lex-SVM. Finally we analyze the properties of Lex-SVM, and conclude with an efficient implementation.

2.1. Exon expression correlation structure

We use the concepts *grouping* and *subgrouping* to describe the correlation structure. A rough idea about exon grouping is that exons of a gene are transcribed together, therefore their expressions are correlated and should be considered as a *group*. However, their correlation might be interrupted by alternative splicing. For more detailed and reliable exon expression correlation structure, we need to take splicing patterns of the genes into consideration. As indicated by Fig. 1, for one gene with multiple splicing forms, the exons could be divided into *subgroups* according to whether they are always present and absent together. Expression of exons in a *subgroup* are expected to be always correlated.

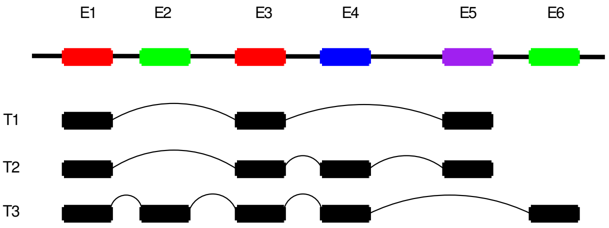

Fig. 1. A gene with 6 exons (labeled as E_1, E_2, E_3, E_4, E_5, E_6), and its 3 transcripts (labeled as T_1, T_2, T_3). The exons can be divided into 4 subgroups according to whether they are always present and absent together: $S_1 = \{E_1, E_3\}, S_2 = \{E_2, E_6\}, S_3 = \{E_4\}, S_4 = \{E_5\}$. Subgroups are marked by different colors.

Such splicing patterns can be obtained from sequence databases like Ensembl [12], which have rapidly accumulated huge amount of transcripts. When this paper is written, sequences of 23,621 protein-coding genes have been deposited in Ensembl. Among them, 11,254 genes have multiple spliced isoforms (total 37,099 transcripts, 3.3 transcripts for each gene).

2.2. Lex-SVM

To incorporate previously described prior knowledge of exon grouping and subgrouping, we propose a novel penalty function in the form of

$$\Omega_{ex}(w) = \sum_{g \in G} \max_{C_g \in S_g} \{ \max_{i,j \in C_g, i \neq j} \{|w_i|, |w_i - w_j|, |w_j|\}\} \tag{3}$$

, where G is the set of genes measured, i and j are the indexes of exons in a gene, S_g is the set of exon subgroups gene g has, and C_g corresponds to an exon subgroup such that $C_g \in S_g$. The design of this penalty is greatly inspired by the *fusion* penalty used in fused lasso[14, 23] and *maximum norm* in F-∞ SVM [31].

Lemma 2.1. *$\Omega_{ex}(w)$ is a norm.*

Proof of this lemma can be found in Appendix. Combining this penalty function with *hinge loss* in standard SVM, we have

$$\min_{w,w_0} \sum_{i=1}^{n} [1 - y_i(\sum_{g \in G} x_{i,(g)}^T w_{(g)} + w_0)]_+ + \lambda \Omega_{ex}(w) \tag{4}$$

, where λ is a tuning parameter. We call the penalty function $\Omega_{ex}(w)$ ex-norm (denoted as $\|w\|_{ex}$) and the above algorithm Lex-SVM. Note that it can be viewed as the Lagrange formulation of the following constrained optimization problem:

$$\begin{aligned} \min_{w,w_0} \quad & \sum_{i=1}^{n} [1 - y_i(\sum_{g \in G} x_{i,(g)}^T w_{(g)} + w_0)]_+ \\ \text{s.t.} \quad & \Omega_{ex}(w) \leq \mu \end{aligned}$$

, where μ is a pre-defined parameter.

Several properties of Lex-SVM are noteworthy. First, Lex-SVM still enjoys the so-called margin maximizing property, which is shown formally in the following theorem:

Theorem 2.1. *Suppose the input data $(x_i, y_i), i = 1, ..., n$ are separable. Let $\hat{w}(\lambda)$ be the solution to problem (4), we have:*

(a) $\lim_{\lambda \to 0} \min_i y_i x_i^T \hat{w}(\lambda) = 1$.

(b) Any convergence point of the normalized solutions $\frac{\hat{w}(\lambda)}{\|\hat{w}(\lambda)\|_{ex}}$ to problem (4) as $\lambda \to 0$ is an ex-norm margin maximizing separating hyper-plane. Consequently, if this hyper-plane is unique, then the solutions converge to it:

$$\lim_{\lambda \to 0} \frac{\hat{w}(\lambda)}{\|\hat{w}(\lambda)\|_{ex}} = \arg\max_{\|w\|_{ex}=1} \min_i y_i x_i^T w$$

The proof of this theorem is similar to that of Rosset et al. [21] and Zou and Yuan [31] (see Appendix). The difference to the margin maximized in standard SVM is that the margin here is measured using ex-norm rather than 2-norm. Note that the following inequalities always hold:

$$\|w\|_\infty \leq \|w\|_2 \leq \|w\|_1; \qquad \|w\|_\infty \leq \|w\|_{ex} \leq \|w\|_1$$

Therefore oftentimes the margin Lex-SVM maximizes is closer to that of standard SVM than both L1-SVM and F-∞ SVM.

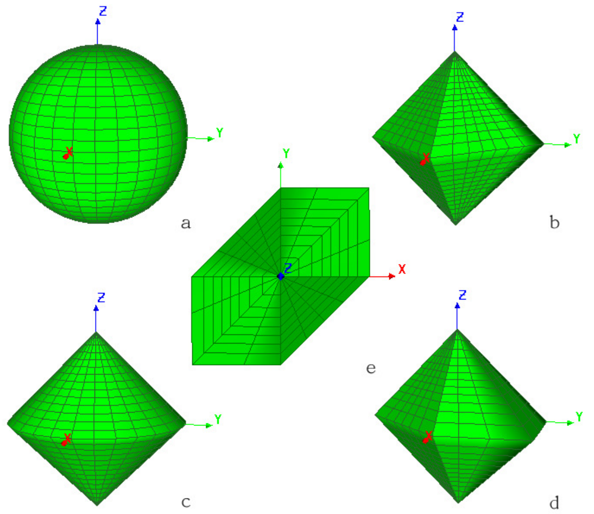

Fig. 2. Equal contours of penalty functions: a) L2, b) Lasso, c) Group Lasso, and d) ex-norm. In this simple problem, the data consists of only three features: x, y and z. In panel c), feature x and y are in the same group, while in panel d), features x and y are in the same subgroup. Panel e) provides a top view of panel d), exposing xy plane.

Second, owing to the singular nature of function *max*, Lex-SVM is able to simultaneously eliminate a group of features. With a proper choice of λ, $\max_{C_g \in S_g}\{\max_{i,j \in C_g, i \neq j}\{|w_i|, |w_i - w_j|, |w_j|\}\}$

will be exactly zero, thus all exons of the gene are excluded.

Third, it enforces *consensus* among exons within a subgroup. We display the feasible solution space of ex-norm and several representative penalty functions in Fig.2. As pointed out in previous studies [5, 29], one can get a good intuition on how penalty functions operate by looking at the singularity of these sets. From Fig.2, we see that *L2 penalty* has no singular points, thus does not favor sparsity. *Lasso*'s singular points lie on the edges of a regular octahedron, thus it selects all variables individually. *Group lasso*'s singular points lie on the two poles of a bicone and $L2$ unit sphere in the xy plane, therefore, it could select features group-wisely but does not prompt similarity within a group.

In contrast, ex-norm has two types of singular points: i) singularity at $w_i = w_j$. Therefore, ex-norm favors $w_i = w_j$ when measurements of exon i and j agree with each other. ii) singularity at $w_i = 0$. This would eliminate exons contradicting the majority within a subgroup or unrelated to the disease. These two attributes are referred as *consensus* property, and are formally presented in the following theorem. Proof can be found in Appendix.

Theorem 2.2. *Given data $(x_i, y_i), i = 1, ..., n$, where the x are standardized. One subgroup contains two features: j and j'. Let $\hat{w}(\lambda)$ be the solution to (4), and $\tilde{w}$ be the solution to the hinge loss $\min_{w,w_0} \sum_{i=1}^{n} [1 - y_i(\sum_{g \in G} x_{i,(g)}^T w_{(g)} + w_0)]_+$.*

(a) if $\tilde{w}_j \tilde{w}_{j'} > 0$, then $\exists \mu > 0$ such that, $\forall \lambda > \mu$, $\hat{w}_j(\lambda) = \hat{w}_{j'}(\lambda)$.

(b) if $\tilde{w}_j \tilde{w}_{j'} < 0$, then $\exists \mu > 0$ such that, $\forall \lambda > \mu$, $\hat{w}_j(\lambda)\hat{w}_{j'}(\lambda) = 0$.

Overall, ex-norm penalization is thus expected to provide solutions with few active groups, encourage features within a subgroup to take equal weights, and exclude features that contradict the majority in the subgroup. In fact, L1-SVM is a particular case of Lex-SVM, where each group only contains one feature. If each subgroup contains only one feature, it can be seen that F-∞ SVM [31] is also a particular case of Lex-SVM.

2.3. Implementation

Introducing slack variables, the problem described in (4) can be solved using linear programming technique:

$$
\begin{array}{rl}
\min\limits_{w,w_0} & \sum\limits_{i=1}^{n} \xi_i + \lambda \sum\limits_{g \in G} m_g \\
\text{s.t.} & \\
\forall i = 1, ..., n & \xi_i \geq 0 \\
 & \xi_i \geq 1 - y_i(x_i^T w + w_0) \\
\forall C_g \in S_g, \forall i \in C_g & w_i \leq m_g \\
 & w_i \geq -m_g \\
\forall i, j \in C_g & w_i - w_j \leq m_g \\
 & w_i - w_j \geq -m_g \\
\forall g \in G & m_g \geq 0
\end{array}
$$

As usual, let $\hat{w}$ and $\hat{w_0}$ be the solution, the fitted classifier can be written as $\hat{f}(x) = x\hat{w} + \hat{w_0}$, and the decision rule is $sign(\hat{f}(x))$.

The time complexity of a linear programming problem is determined by the number of variables and the number of constraints. Lex-SVM uses the same number $(n + |G|)$ of variables as F-∞ SVM, but much fewer than L1-SVM as one auxiliary variable has to be defined for each w_i in L1-SVM. The major complexity raised in Lex-SVM is the new constraints added between each two exons in a subgroup. Luckily, the size of a subgroup can be bounded by a constant (e.g. 100) and Megiddo [17] proved that linear programming problem can be solved in linear time when the dimension is fixed, therefore Lex-SVM share the same time complexity as L1-SVM and F-∞ SVM. However, in practice as the linear programming is usually solved using the simplex or prime-dual method, the actual time costs may vary depending on the property of constraint matriex.

3. RESULTS

Roughly speaking, exon expression profile contains two types of information, namely, gene expression information and alternative splicing information. Although the association between splicing and disease is frequently reported [26], to what extent this association help disease classification is still unclear. To investigate that, we first used L1-SVM and SVM-RFE [10] to classify gene expression profile, splicing profile (containing only splicing information, see below), and exon expression profile, where gene abundance and splicing information are coupled together. Then, we used Lex-SVM on exon expression profiles and compared it with L1-SVM and SVM-RFE. Before listing the test results, it is worth mentioning that SVM-RFE is another popular feature selection method. Quite different from regularization approaches like L1-SVM, SVM-RFE performs feature selection outside the training process. It trains a SVM classifier iteratively with the current set of features and ranks the features according to their weights in the trained SVM. In each iteration, one or several features that rank lowest in cross validation are removed.

3.1. Data set

In our experiments, we used four Affymetrix exon array datasets representing different kinds of applications, different sample sizes and different sample preparation methods. The first one is a publicly available human colon cancer data including 10 tumor/normal paired specimens [8]. The second one is a human gastric caner dataset containing 50 tumor/normal paired specimens (GEO Accession #: GSE13195). The third one is a brain cancer data including 26 glioblastomas and 23 oligodendrogliomas brain samples (GEO Accession #: GSE9385) [6]. The last one is a human cancer datasets containing 84 breast tumor and 43 lung cancer samples (GEO Accession #: GSE16534) [15]. On the first two datasets, the objective is to distinguish cancerous tissues from normal ones; while on the last two datasets, the objective is to distinguish different types of cancers.

These exon array datasets were preprocessed as follows: First, probeset signals were estimated using PLIER (with GC-normalization option) to generate exon expression profile [1]. For each sample, another two derivative profiles have been constructed for comparison: i) gene expression profile. We summarized the intensity of each gene's meta-probe set using PLIER. ii) splice profile. To evaluate the contribution of alternative splicing alone to classification, we calculated a splice profile through $\log_2 \frac{E}{G}$, where E is the exon signal and G is the corresponding gene signal. This normalized intensity has been widely used in other studies to quantify splicing [2, 8]. Second, for efficiency of the algorithms, we required the genes to be significantly expressed in at least one of the comparison groups, leaving 2,579 Ensembl genes (38,156 exon probesets) in colon cancer dataset, 3,248 genes (59,236 exon probesets) in gastric cancer dataset, 2,922 genes (47,967 exon probesets) in brain tumor dataset, and 3,759 genes (62,453 exon probesets) in breast/lung cancer dataset. Finally, all features (exon signal, gene signal or splicing intensity) are scaled across training and testing samples to zero mean and unit variance. The objective is to avoid features in greater numeric ranges dominate those in smaller numeric ranges.

3.2. Splicing information improves classification

As far as we know, till now there is no report on using exon array profile for disease classification. To estimate the contribution of splicing information to disease classification, we first used L1-SVM and SVM-RFE to classify gene expression profile, splicing profile and exon expression profile, respectively. For each dataset, we performed 10-fold cross-validation (CV) to select the best parameters and evaluate the algorithms' performance. For L1-SVM, we trained it for wide-ranging values of the parameter λ. For SVM-RFE [10], at step n bottom $\frac{1}{\min\{n+1,10\}}$ of the remaining features were removed to expedite the selection procedure. The minimum CV errors are adjusted for bias according to Tibshirani and Tibshirani [25], and are listed in Table 1.

The results show that using splicing information alone could achieve comparable classification power as gene abundance. The underlying reason might be the reciprocal cause-effect relationship between aberrations in alternative splicing and aberrations in transcription [26]. Compared with gene expression profile, exon expression profile gives rise to an improvement of $\sim$ 5% in predictive power. This finding is consistent with the previous observation by Zhang et al. [28]. They used SVM-RFE to compare the classification power of isoform and gene expression profiles, and also found an increase of $\sim$ 5%. Oftentimes, SVM-RFE is more accurate than L1-SVM, but it also requires $\sim$30 more features (see Table 1).

On exon expression profile, we observed that with few exceptions L1-SVM selected only one exon

Table 1. Error rates of L1-SVM and SVM-RFE in 10-fold CV. The numbers in bracket are the avarage sizes of features selected in CV.

	Colon		Gastric		Brain		Breast/lung	
	L1	RFE	L1	RFE	L1	RFE	L1	RFE
Gene profile	0.35 (13)	0.30 (37)	0.11 (32)	0.10 (65)	0.20 (29)	0.22 (41)	0.02 (31)	0.02 (46)
Splice profile	0.35 (14)	0.35 (56)	0.14 (36)	0.10 (75)	0.30 (34)	0.25 (72)	0.04 (35)	0.02 (55)
Exon profile	0.30 (15)	0.25 (47)	0.10 (38)	0.06 (75)	0.15 (31)	0.15 (54)	0.02 (35)	0 (49)

in a gene. For instance, under optimal value of parameter λ we observed that 16 exons were selected from 16 different genes on colon cancer dataset, and 32 exons were selected from 31 different genes on brain tumor dataset. This might cause two problems: first, signals of a single exon are likely to be influenced by noise, limiting the accuracy of L1-SVM in 10-fold CV; second, one exon does not tell us much about which isoform of the gene is associated with the disease. See Fig. 1 for illustration, if only exon E_5 is selected, we could not discriminate isoforms T_1 and T_2, unless exon E_4 is also selected. Due to these problems, the increase in classification power from gene to exon expression profile might be underestimated. In subsequent subsection, we investigated the performance of Lex-SVM, where prior knowledge of exon expression correlation structure is incorporated to tackle these problems.

3.3. Performance of Lex-SVM

We tested Lex-SVM on the four exon expression profiles under a wide range values of parameter λ. Besides L1-SVM and SVM-RFE, fused SVM [23, 19] has also been tested for comparison. In fused SVM, the fusion penalty was applied on the exons of a gene sequentially. When all four methods are under their optimal parameter settings, Lex-SVM is over 5% more accurate than L1-SVM on all four datasets, about 5% more accurate than SVM-RFE on gastric and brain datasets, and over 2% more accurate than fused SVM on three datasets (Table 2). Lex-SVM is also more accurate than the other three methods under most non-optimal parameter settings (data not shown).

Here, we consider a gene *selected* if it has at least one exon with non-zero weight. Under optimal value of parameter λ, ∼ 12 genes were selected by Lex-SVM on colon cancer dataset, ∼ 20 genes on brain tumor dataset, ∼ 25 on gastric cancer dataset, and ∼ 27 on breast/lung cancer dataset. The frequently-selected genes (over 8 times in 10-fold CV) in colon cancer data are *PRKDC, PRPF8, ITGB4,* and *VWF*. Among them *ITGB8* has been validated by RT-PCR to be differentially spliced in the comparison groups [8]. In brain tumor data the genes are *VWF, PRKDC, LRP1, PREX1, MACF1, PDZD2, CHL1,* and *ATP2B4*. Among them, *ATP2B4* has been validated by RT-PCR to be differentially spliced [6]. All of them were previously reported in these diseases (Table 3). Limited by the space, the genes frequently selected in gastric and breast/lung cancer datasets will not be listed.

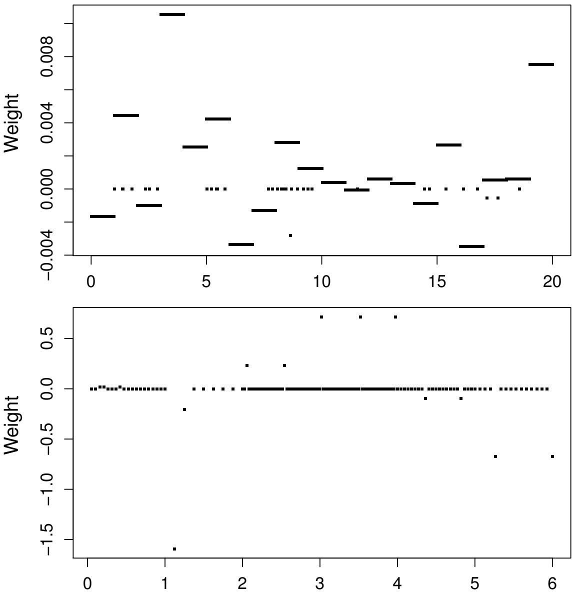

Fig. 3. The weights of the exons selected by Lex-SVM on brain tumor dataset when $log_2\lambda = 0$ (top panel) and $log_2\lambda = -40$ (bottom panel). Each unit on x axis represents one gene, and the weights of its exons are drawn within this unit sequentially. Points with weight≠0 are exons selected by Lex-SVM.

According to the design of Lex-SVM, exons of the selected genes would be assigned equal weights when parameter λ is large enough. The weights of exons in the selected genes are shown in Fig. 3. It

Table 2. Performance of Lex-SVM, L1-SVM, fused SVM and SVM-RFE in 10-fold CV. The first row shows the error rate in CV, the second row shows the average number of genes selected in each round of CV, and the last row shows the number of genes selected over five times in CV.

Colon				Gastric				Brain				Breast/lung			
Lex	L1	fused	RFE	Lex	L1	fused	RFE	Lex	L1	fused	RFE	Lex	L1	fused	RFE
0.25	0.30	0.25	0.25	0.02	0.1	0.16	0.06	0.08	0.15	0.1	0.15	0	0.02	0.02	0
12	15	14	47	25	38	32	75	20	31	25	54	27	35	35	50
10	6	9	-	20	21	22	-	19	27	23	-	26	16	26	-

can be seen that when $log_2\lambda = 0$ (optimal value of λ), Lex-SVM forces most of the exons in a selected gene to take equal weight (forming a series of horizontal lines in Fig. 3), meanwhile excluds the rest exons that contradict the majority in their subgroups or unrelated to the disease (scattering points along $weight = 0$ in Fig. 3). When $log_2\lambda = -40$ (less-than optimum), Lex-SVM only selects several exons from a gene. This confirms the statements in Theorem 2.2 that a large-enough λ is required to guarantee the consensus property.

In Table 2, we compared the number of genes selected by Lex-SVM, L1-SVM, fused SVM and SVM-RFE. It can be seen that Lex-SVM is able to generate a more compact model (less genes) than L1-SVM, SVM-RFE and fused SVM does. More importantly, Lex-SVM selects genes more consistently than L1-SVM and fused SVM from round to round in 10-fold CV. Among those genes selected by Lex-SVM, over 80% were selected more than five times in 10-fold CV, while the proportion in L1-SVM is less than 50% and in fused SVM about 70%. It should be noted that in each step SVM-RFE only eliminates one or several features from last subset, therefore the genes it selects will not change much from round to round in CV.

Unlike L1-SVM selecting only one exon from a gene, Lex-SVM assigns equal weights to as many exons of a gene as possible, therefore facilitating further determination of which isoform is associated with the disease. According to the exons selected, we can narrow the plenty of transcripts recorded in Ensembl down to $1 \sim 3$ candidates for each gene. Limited by the space, in Table 3 we only listed the genes selected for colon cancer and brain tumor dataset. For example, RT-PCR found that ITGB4 tends to skip exon 35 in colon cancer tissues [8]. In our experiments, Lex-SVM also excluded this exon and the corresponding transcript ENST00000200181 can be determined from the exons selected.

4. DISCUSSION

The usage of exon expression profiles, which contains both transcription and splicing information, gave popular classification methods like L1-SVM and SVM-RFE only a modest increase (5%) in accuracy. The underlying reasons of this limited increase might be: i) On our testing datasets, using gene abundance alone can reach a disease classification accuracy of 80%, leaving an improvement space of only 20%. For diseases that are caused by aberrations in splicing, simply using gene abundance might not achieve such high accuracy as in our experiments, then the importance of introducing splicing information will be more apparent.

ii) Existing approaches like L1-SVM and SVM-RFE cannot fully utilize the abundant information embedded in exon expression profiles due to one common shortage: neglecting the intrinsic correlation structure in exon expression profile. Feature selection approaches for SVM fall into three categories [9]: filter (e.g. t-test), wrapper (e.g. SVM-RFE) and embedded methods (e.g. L1-SVM). Several algorithms in the last category such as group lasso [27] and fused lasso [23] have considered feature correlation structure in the selection process, but it is unclear how feature correlation structure can be addressed in filter and wrapper approaches.

Lex-SVM is also an embedded method. It uses a particular regularization term (ex-norm) to translate our prior knowledge on exon expression correlation into constraints on the classifier. Lex-SVM is able to select a restricted number of genes and assign equal weights to as many exons of a gene as possible. The sample size of exon expression profiles currently deposited in GEO ranges from a few to dozens. Tests on such small size datasets demonstrate that Lex-SVM generally performs better than classical tech-

Table 3. Details of the genes selected over 8 times.

Colon cancer			
Gene	#Transcripts[a]	Narrowed transcripts[b]	PMID[c]
PRKDC	2	ENST00000314191	7624361
PRPF8	1	ENST00000304992	15078890
ITGB4	5	ENST00000200181	17192196
VWF	3	ENST00000453974, ENST00000321023, ENST00000261405	15810086, 16254106

Brain tumor			
Gene	#Transcripts[a]	Narrowed transcripts[b]	PMID[c]
VWF	3	ENST00000453974, ENST00000321023, ENST00000261405	2245394
PRKDC	2	ENST00000314191	10629611, 19435898
LRP1	3	ENST00000393833, ENST00000338962, ENST00000243077	9879460, 9205092
PREX1	5	ENST00000371941	15273287
MACF1	17	ENST00000404645, ENST00000361689	15803373
PDZD2	4	ENST00000282493	11519051
CHL1	16	ENST00000397491, ENST00000256509	10103075
ATP2B4	9	ENST00000357681, ENST00000341360	17575129

[a] the number of transcripts recorded in Ensembl.
[b] related transcripts determined from selected exons. Multiple transcripts will be listed if the selected exons can not discriminate them.
[c] PubMed ID of the supporting literatures.

niques that do not take into account the specificities of exon expression profile.

Our method is closely related to F-∞ SVM and fused SVM [19, 24]. These two techniques have accomplished groupwise feature selection and local constancy of the coefficient profile respectively. However if directly applied to exon expression profile, F-∞ SVM could not garantee that the exons in a selected group have similar weights. The constancy property of fused lasso used in fused SVM is limited on a local region (adjacent features), therefore could not enforce two exons that are always expressed together but not adjacent in sequence to have similar weights. Also due to different genes have various numbers of exons, direct summation of fusion penalty as in fused SVM can cause higher penalty on genes with more exons. Our major contribution is embedding fused lasso in a maximum norm penalty to meet the special requirements of exon expression profile classification. This modification extends the local constancy of fused lasso to groupwise constancy, and treats genes with different numbers of exons equally.

There is one pitfall in our model: the splicing patterns in Ensembl are not either complete or specially deposited for the disease under study. If the disease involves new alternative splicing events that are not recorded in Ensembl, our regularization approach might unfavorably encourage two differentially spliced exon to have similar weights. To avoid such pitfall, the users can first run alternative splicing detecting methods to see whether novel splicing event exists, or simply divides the exons of the gene each in different subgroups.

Lex-SVM can also be used to process RNA-Seq data. Exon expression profile from RNA-Seq data can be generated through calculating RPKM, i.e., normalizing the counts of reads mapped to an exon against the exon length and million mapped reads to the transcriptome [18]. Simply due to the unavailability of such public dataset, we could not report Lex-SVM's performance on RNA-Seq data in this paper.

An interesting observation from Fig. 3 is that as tuning parameter λ diminishes to less-than optimal value and ex-norm gradually loses its consensus property, the number of genes selected first decreases and then increases rather than increase directly. To deeply understand this phenomena, the whole solution path (changes of the solution with λ) of Lex-SVM is needed. Luckily, Lex-SVM is piecewise linear, therefore theoretically we can generate the whole regularized path simply by calculating the 'step sizes' between each two consecutive 'joints' [20, 11]. Developing an efficient algorithm to calculate the whole solution path would be our future work.

Appendix

Proof of lemma 2.1

Proof. Positive homogeneity holds trivially; thus, it suffices to prove the positive definiteness and triangle inequality properties.

—positive definiteness: Note that $\Omega_{ex}(w) \geq \sum_{g \in G} \max_{i \in E_g}\{|w_i|\}$, where E_g are the exons gene g contains. If $\Omega_{ex}(w) = 0$, we have $\max_{i \in E_g}\{|w_i|\} = 0$ for any g, meaning $w = 0$. Conversely, we have $\Omega_{ex}(w) = 0$ when $w = 0$. Therefore, $\Omega_{ex}(w)$ is positive definite.

—triangle inequality: Considering two weight vectors w and w', we have $\forall i, |w_i + w'_i| \leq |w_i| + |w'_i|$, and $\forall i, j, |w_i - w_j + w'_i - w'_j| \leq |w_i - w_j| + |w'_i - w'_j|$; thus, $\Omega_{ex}(w + w') \leq \Omega_{ex}(w) + \Omega_{ex}(w')$.

In fact, $\Omega_{ex}(w)$ can be viewed as L_∞ of an expanded vector $\tilde{w}$. Besides the elements of w, $\tilde{w}$ also contains absolute difference between some elements in w. □

Proof of theorem 2.1

Proof. Part (a) We first show that $\liminf_{\lambda \to 0}\{\min_i y_i x_i^T \hat{w}(\lambda)\} \geq 1$. Assume the contrary: there is a decreasing sequence of $\{\lambda_k\} \to 0$ and some $\epsilon > 0$ such that, for all k, $\min_i y_i x_i^T \hat{w}(\lambda_k) \leq 1 - \epsilon$. Then $L(\hat{w}(\lambda_k), \lambda_k) \geq [1 - (1 - \epsilon)]_+ = \epsilon$. Let $m_0 = \min_i y_i x_i^T w_0 > 0$ and $w' = \frac{w_0}{m_0}$. As $\hat{w}(\lambda_k) = \arg\min_w L(w, \lambda_k)$, we have $L(w', \lambda_k) \geq L(\hat{w}(\lambda_k), \lambda_k) \geq \epsilon$. However, note that $\min_i y_i x_i^T w' = 1$, therefore $L(w', \lambda_k) = \lambda_k \sum_{g \in G} \Omega_{ex}(w') \to 0$ as $\lambda_k \to 0$. This contradicts with $L(w', \lambda_k) \geq \epsilon$.

To show $\limsup_{\lambda \to 0}\{\min_i y_i x_i^T \hat{w}(\lambda)\} \leq 1$, similarly we assume the contrary: there is a decreasing sequence of $\{\lambda_k\} \to 0$ and some $\epsilon > 0$ such that, for all k, $\min_i y_i x_i^T \hat{w}(\lambda_k) \geq 1 + \epsilon$. Then, $L(\hat{w}(\lambda_k), \lambda_k) = \lambda_k \Omega_{ex}(\hat{w}(\lambda_k))$, and $L(\frac{\hat{w}(\lambda_k)}{1+\epsilon}, \lambda_k) = \frac{\lambda_k}{1+\epsilon}\Omega_{ex}(\hat{w}(\lambda_k))$. So $L(\frac{\hat{w}(\lambda_k)}{1+\epsilon}, \lambda_k) < L(\hat{w}(\lambda_k), \lambda_k)$, which contradicts the definition of $\hat{w}(\lambda_k)$. In conclusion, $\lim_{\lambda \to 0} \min_i y_i x_i^T \hat{w}(\lambda) = 1$.

Part (b) Suppose a subsequence of $\frac{\hat{w}(\lambda_k)}{\|\hat{w}(\lambda_k)\|_{ex}}$ converges to w^* as $\lambda_k \to 0$. Then $\|w^*\|_{ex} = 1$. Denote $\min_i y_i x_i^T w$ by $m(w)$. We need to show that $m(w^*) = \max_{\|w\|_{ex}=1} m(w)$. Assume the contrary: there is some w' such that $\|w'\|_{ex} = 1$ and $m(w') > m(w^*)$. From part(a), we have

$$\begin{aligned}
&\lim_{\lambda_k \to 0} m(w^*)\|\hat{w}(\lambda_k)\|_{ex} \\
=\ &\lim_{\lambda_k \to 0} \min_i y_i x_i^T \frac{\hat{w}(\lambda_k)}{\|\hat{w}(\lambda_k)\|_{ex}}\|\hat{w}(\lambda_k)\|_{ex} \\
=\ &1
\end{aligned}$$

. On the other hand, we have

$$L(\frac{w'}{m(w')}, \lambda_k) = \lambda_k \|\frac{w'}{m(w')}\|_{ex} = \frac{\lambda_k}{m(w')}$$

$$L(\hat{w}(\lambda_k), \lambda_k) \geq \lambda_k \|\hat{w}(\lambda_k)\|_{ex}$$

Thus,

$$\frac{L(\frac{w'}{m(w')}, \lambda_k)}{L(\hat{w}(\lambda_k), \lambda_k)} \leq \frac{m(w^*)}{m(w')} \frac{1}{m(w^*)\|\hat{w}(\lambda_k)\|_{ex}}$$

$$\lim_{\lambda_k \to 0} \sup \frac{L(\frac{w'}{m(w')}, \lambda_k)}{\text{Ł}(\hat{w}(\lambda_k), \lambda_k)} \leq \frac{m(w^*)}{m(w')} < 1$$

which contradicts the definition of $\hat{w}(\lambda_k)$. Therefore, w^* is an ex-norm margin maximizing separating hyper-plane. □

Proof of theorem 2.2

Proof. Part (a) If $\tilde{w}_j \tilde{w}_{j'} > 0$, then $\hat{w}_j(\lambda)\hat{w}_{j'}(\lambda) \geq 0$ because opposite sign increases not only hinge loss but also ex-norm penalty. To show that $\hat{w}_j = \hat{w}_{j'}$, assume the contrary: $\hat{w}_j \neq \hat{w}_{j'}$. If $|\hat{w}_j| > |\hat{w}_{j'}|$, consider w^* as follows:

$$w^*_k = \begin{cases} \hat{w}_k & \text{if } k \neq j \text{ and } k \neq j' \\ \frac{1}{2}(\hat{w}_j + \hat{w}_{j'}) & \text{if } k = j \text{ or } k = j' \end{cases}$$

Thus,

$$\begin{aligned}
&L(\hat{w}, \lambda) - L(w^*, \lambda) \\
=&\sum_i \{[1 - y_i(\sum_{g \in G} x_{i,(g)}^T \hat{w}_{(g)} + \hat{w}_0)]_+ \\
&- [1 - y_i(\sum_{g \in G} x_{i,(g)}^T w^*_{(g)} + w^*_0)]_+\} + \lambda\{\Omega(\hat{w}) - \Omega(w^*)\} \\
\geq& - \sum_i |x_{i,j}\hat{w}_j + x_{i,j'}\hat{w}_{j'} - \frac{1}{2}(x_{i,j} + x_{i,j'})(\hat{w}_j + \hat{w}_{j'})| \\
&+ \frac{\lambda}{2}|\hat{w}_j - \hat{w}_{j'}|
\end{aligned}$$

For any $\lambda > \frac{2}{|\hat{w}_j - \hat{w}_{j'}|}\sum_i |x_{i,j}\hat{w}_j + x_{i,j'}\hat{w}_{j'} - \frac{1}{2}(x_{i,j} + x_{i,j'})(\hat{w}_j + \hat{w}_{j'})|$, we have $L(\hat{w}, \lambda) - L(w^*, \lambda) > 0$. This contradicts with the definition of $\hat{w}$. Therefore, $\exists \mu > 0$ such that, $\forall \lambda > \mu$, $\hat{w}_j(\lambda) = \hat{w}_{j'}(\lambda)$.

Part (b) To show that $\hat{w}_j \hat{w}_{j'} = 0$, assume the contrary: $\hat{w}_j \hat{w}_{j'} \neq 0$. If $|\hat{w}_j| > |\hat{w}_{j'}|$, consider w^* as follows:

$$w^*_k = \begin{cases} \hat{w}_k & \text{if } k \neq j \text{ and } k \neq j' \\ \frac{1}{2}(\hat{w}_j - \hat{w}_{j'}) & \text{if } k = j \text{ or } k = j' \end{cases}$$

Thus,

$$\begin{aligned} & L(\hat{w}, \lambda) - L(w^*, \lambda) \\ = & \sum_i \{[1 - y_i(\sum_{g \in G} x^T_{i,(g)} \hat{w}_{(g)} + \hat{w}_0)]_+ \\ & - [1 - y_i(\sum_{g \in G} x^T_{i,(g)} w^*_{(g)} + w^*_0)]_+\} + \lambda\{\Omega(\hat{w}) - \Omega(w^*)\} \\ \geq & - \sum_i |x_{i,j}\hat{w}_j + x_{i,j'}\hat{w}_{j'} - \frac{1}{2}(x_{i,j} + x_{i,j'})(\hat{w}_j - \hat{w}_{j'})| \\ & + \frac{\lambda}{2}(|\hat{w}_j| + |\hat{w}_{j'}|) \end{aligned}$$

For any $\lambda > \frac{2}{|\hat{w}_j| + |\hat{w}_{j'}|} \sum_i |x_{i,j}\hat{w}_j + x_{i,j'}\hat{w}_{j'} - \frac{1}{2}(x_{i,j} + x_{i,j'})(\hat{w}_j - \hat{w}_{j'})|$, we have $L(\hat{w}, \lambda) - L(w^*, \lambda) > 0$. This contradicts with the definition of $\hat{w}$. Therefore, $\exists \mu > 0$ such that, $\forall \lambda > \mu$, $\hat{w}_j(\lambda)\hat{w}_{j'}(\lambda) = 0$. □

References

1. Affymetrix. Guide to probe logarithmic intensity error (plier) estimation. Technical report, 2005.
2. Affymetrix. Identifying and validating alternative splicing events. Technical report, Affymetrix, 2006.
3. F. Bartel, H. Taubert, and L. C. Harris. Alternative and aberrant splicing of MDM2 mRNA in human cancer. *Cancer Cell*, 2(1):9–15, 2002.
4. D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, 72:291–336, 2003.
5. Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, December 2001.
6. Pim J French, Justine Peeters, Sebastiaan Horsman, Elza Duijm, Ivar Siccama, Martin J van den Bent, Theo M Luider, Johan M Kros, Peter van der Spek, and Peter A Sillevis Smitt. Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res*, 67(12):5635–5642, Jun 2007.
7. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, Oct 2000.
8. P. J. Gardina, T. A. Clark, B. Shimada, M. K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee, C. Davies, A. Williams, and Y. Turpaz. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, 7:325, 2006.
9. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.
10. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, 2002.
11. Trevor Hastie, Saharon Rosset, Robert Tibshirani, Ji Zhu, and Nello Cristianini. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.
12. T. J P Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Res*, 37(Database issue):D690–D697, Jan 2009.
13. Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-wide survey of human alternative pre-mrna splicing with exon junction microarrays. *Science*, 302(5653):2141–2144, Dec 2003.
14. S. R. Land and J. H. Friedman. Variable fusion: A new adaptive signal regression method. Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh, 1997.
15. Eva Lin, Li Li, Yinghui Guan, Robert Soriano, Celina Sanchez Rivers, Sankar Mohan, Ajay Pandita, Jerry Tang, and Zora Modrusan. Exon array profiling detects eml4-alk fusion in breast, colorectal, and non-small cell lung cancers. *Mol Cancer Res*, 7(9):1466–1476, Sep 2009.
16. J. Lukas, D. Q. Gao, M. Keshmeshian, W. H. Wen, D. Tsao-Wei, S. Rosenberg, and M. F. Press. Alternative and aberrant messenger RNA splicing of the mdm2 oncogene in invasive breast cancer. *Cancer Res*, 61(7):3212–9, 2001.
17. Nimrod Megiddo. Linear programming in linear time when the dimension is fixed. *J. ACM*, 31(1):114–127, 1984.
18. Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq.

Nat Methods, 5(7):621–628, Jul 2008.

19. Franck Rapaport, Emmanuel Barillot, and Jean-Philippe Vert. Classification of arraycgh data using fused svm. *Bioinformatics*, 24(13):i375–i382, Jul 2008.
20. Saharon Rosset and Ji Zhu. Title: Piecewise linear regularized solution paths. *Annals of Statistics*, 35:1012–1030, 2007.
21. Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *In Advances in Neural Information Processing Systems (NIPS) 15*, page 16. MIT Press, 2003.
22. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
23. Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal Of The Royal Statistical Society Series B*, 67(1):91–108, 2005.
24. Robert Tibshirani and Pei Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 9(1):18–29, Jan 2008.
25. Ryan J. Tibshirani and Robert Tibshirani. A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics*, 3(2):822–829, 2009.
26. Guey-Shin Wang and Thomas A Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet*, 8(10):749–761, Oct 2007.
27. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.
28. Chaolin Zhang, Hai-Ri Li, Jian-Bing Fan, Jessica Wang-Rodriguez, Tracy Downs, Xiang-Dong Fu, and Michael Q Zhang. Profiling alternatively spliced mrna isoforms for prostate cancer classification. *BMC Bioinformatics*, 7:202, 2006.
29. Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37:3468–3497, 2009.
30. Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems*, page 16. MIT Press, 2003.
31. Hui Zou and Ming Yuan. The f_∞-norm support vector machine. *Statistica Sinica*, 18:379–398, 2008.