FINDING NEEDLES IN COMPRESSED HAYSTACKS

Robert Calderbank and Sina Jafarpour

ABSTRACT

In this paper, we investigate the problem of compressed learning, i.e. learning directly in the compressed domain. In particular, we provide tight bounds demonstrating that the linear kernel SVMs classifier in the measurement domain, with high probability, has true accuracy close to the accuracy of the best linear threshold classifier in the data domain. Furthermore, we indicate that for a family of well-known deterministic compressed sensing matrices, compressed learning is provided on the fly. Finally, we support our claims with experimental results in the texture analysis application.

Index Terms— Compressed Learning, Support Vector Machines, Delsarte-Goethals Frames, Texture Analysis.

1. INTRODUCTION

In many applications, the data has a sparse representation in some basis in a much higher dimensional space. Examples are the sparse representation of images in the wavelet domain, the bag of words model of text, and the routing tables in data monitoring systems.

Compressed sensing combines measurement to reduce the dimensionality of the underlying data with reconstruction to recover sparse data from the projection in the measurement domain. However there are many sensing applications where the objective is not full reconstruction but is instead classification with respect to some signature. Examples include radar, detection of trace chemicals, face detection and video streaming [1] where we might be interested in anomalies corresponding to changes in wavelet coefficients in the data domain. In all these cases our objective is pattern recognition in the measurement domain.

Classification in the measurement domain offers a way to resolve this challenge and we show that it is possible to design measurements for which there are performance guarantees. Similar to compressed sensing, *linear* measurements are used to remove the costs of pointwise sampling and compression. However, the ultimate goal of compressed learning is not reconstruction of the sparse data from their linear measurements. In contrast, here we are provided with compressively sampled training data, and the goal is to design a classifier directly in the measurement domain with almost the same accuracy as the best classifier in the data domain.

Being able to learn in the compressed domain is beneficial both from the compressed sensing and the machine learning points of view. From the compressed sensing view-point, it eliminates the significant cost of recovering irrelevant data; in other words, classification in the measurement domain is like a sieve and makes it possible to only recover the desired signals, or even remove the recovery phase totally, if we are only interested in the results of classification. This is like finding a needle in a compressively sampled haystack without recovering all the hay. In addition, compressed learning has the potential of working successfully in situations where one can not observe the data domain or where measurements are difficult or expensive.

Dimensionality reduction is a fundamental step in applications as diverse as the nearest-neighbor approximation [2], data-streaming [3], machine learning [1], graph approximation [4], etc. In compressed learning, the sensing procedure can be also considered as a linear dimensionality reduction step. In this paper we will show that most compressed sensing matrices also provide the desired properties of good linear dimensionality reduction matrices.

In terms of geometry, the difference between compressed sensing and compressed learning is that the former is concerned with separating pairs of points in the measurement domain to enable unambiguous recovery of sparse data while the latter is concerned with consistent separation of clouds of points in the data and measurement domains. In this paper we demonstrate feasibility of pattern recognition in the measurement domain. We provide PAC-style bounds guaranteeing that if the data is measured directly in the compressed domain, a soft margin SVM classifier that is trained based on the compressed data performs almost as well as the best possible SVM classifier in the data domain. The result are robust against the noise in the measurement.

The compressed learning framework is applicable to any sparse high-dimensional dataset. For instance, in texture analysis [5] the goal is to predict the direction of an image by looking only at its wavelet coefficients. A weighted voting among horizontal and vertical wavelet coefficients of each image can accurately predict whether the image is vertical, horizontal, or neither. However, in compressive imaging,

S.J. is with the Multimedia Research Group, Yahoo! Research, email: *sina2jp@yahoo-inc.com*. R.C. is with the Department of Computer Science, Duke University, email: *calderbk@cs.duke.edu*. The work is done while S. J. was a Ph.D. student in Princeton University. The work is supported in part by NSF under grant DMS 0701226, by ONR under grant N00173-06-1-G006, and by AFOSR under grant FA9550-09-1-0551.

the high-dimensional wavelet representation of the image is not provided. In contrast, a non-adaptive low-rank sensing matrix is used to project the wavelet vector into some low-dimensional space. Here we show that a weighted voting among the entries of these measurement vectors has approximately the same accuracy as the original weighted voting among the wavelet entries in the texture analysis task.

2. BACKGROUND AND NOTATION

Let n be a positive integer, and let k be a positive integer less than n. We sometimes denote $\{1, \cdots, n\}$ by $[n]$.

We assume that all data are represented as vectors in $\mathbb{R}^n$. The feature space $\mathcal{X}$ (which we also call the data-domain), is a subset of the whole n-dimensional space, and every data point is a vector in the feature space.

Let A be an $m \times n$ matrix. We use the notation A_j for the j^{th} column of the sensing matrix A; its entries will be denoted by a_{ij}, with the row label i varying from 0 to $m-1$. The matrix A is a tight-frame with redundancy $\frac{n}{m}$ if and only if $AA^\dagger = \frac{n}{m}\mathrm{I}_{m\times m}$. Note that if A is a tight-frame with redundancy $\frac{n}{m}$, then $\|A\|^2 = \frac{n}{m}$.

3. SUPPORT VECTOR MACHINES

A support vector machines (SVM) [6] is a linear threshold classifier in some feature space, with maximum margin and consistent with the training examples. Any linear threshold classifier $w(\boldsymbol{x})$ corresponds to a vector $\boldsymbol{w} \in \mathbb{R}^n$ such that $w(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^\top \boldsymbol{x}\right)$; as a result, we identify the linear threshold classifiers with their corresponding vectors. Also for simplicity we only focus on classifiers passing through the origin. The results can be simply extended to the general case.

Whenever the training examples are not linearly separable soft margin SVM are used. The idea is to simultaneously maximize the margin and minimize the empirical hinge loss. More precisely let

$$H(x) \doteq (1+x)_+ = \max\{0, 1+x\},$$

and let $S \doteq \langle(\boldsymbol{x_1}, \mathrm{l}_1), \cdots, (\boldsymbol{x_M}, \mathrm{l}_M)\rangle$ be a set of M labeled training data sampled i.i.d from some distribution $\mathcal{D}$. For any linear classifier $\boldsymbol{w} \in \mathbb{R}^n$ we define its true hinge loss as

$$H_D(\boldsymbol{w}) \doteq \mathbb{E}_{(\boldsymbol{x},\mathrm{l})\sim\mathcal{D}}\left[\left(1 - \mathrm{l}\boldsymbol{w}^\top \boldsymbol{x}\right)_+\right],$$

and its empirical hinge loss

$$\hat{H}_S(\boldsymbol{w}) \doteq \mathbb{E}_{(\boldsymbol{x}_i,\mathrm{l}_i)\sim S}\left[\left(1 - \mathrm{l}_i\boldsymbol{w}^\top \boldsymbol{x}_i\right)_+\right].$$

4. EXPLICIT MATRICES WITH AVERAGE-CASE JOHNSON-LINDENSTRAUSS PROPERTY

4.1. Global Measures of Coherence

Let A be an $m \times n$ matrix such that every column of A has unit ℓ_2 norm. The following two quantities measure the coherence between the columns of A [7]:

- Worst-case coherence $\mu \doteq \max_{\substack{i,j\in[n]\\ i\neq j}} \left|{A_i}^\top A_j\right|$.
- Average coherence $\nu \doteq \frac{1}{n-1}\max_{i\in[n]} \left|\sum_{\substack{j\in[n]\\ j\neq i}} {A_i}^\top A_j\right|$.

Roughly speaking, we can consider the columns of A as n distinct points on the unit sphere in $\mathbb{R}^n$. Worst-case coherence then measures how close two distinct points can be, whereas the average coherence is a measure of the spread of these points.

4.2. Average-Case Distance-Preserving using Delsarte-Goethals Frames

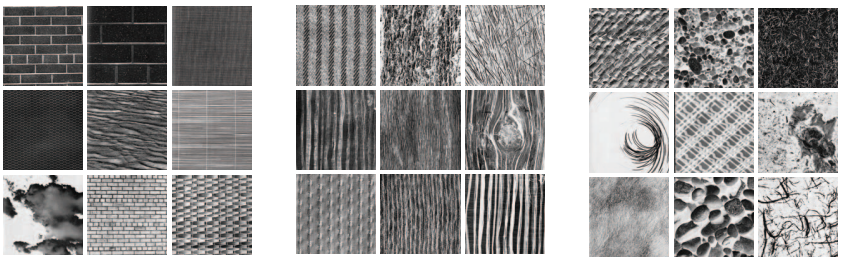
In the previous section, we introduced two fundamental measures of coherence between the columns of a tight-frame. In this section we construct an explicit sensing matrix (*Delsarte-Goethals frame* [8]) with sufficiently small average coherence ν, and worst-case coherence μ, and show how this coherence optimality can be related to the performance of the SVM classifier in the measurement domain.

We start by picking an odd number o. The 2^o rows of the Delsarte-Goethals frame A are indexed by the binary o-tuples t, and the $2^{(r+2)o}$ columns are indexed by the pairs (P, b), where P is an $o \times o$ binary symmetric matrix in the Delsarte-Goethals set $DG(o, r)$ [8], and b is a binary o-tuple. The entry $a_{(P,b),t}$ is given by

$$a_{(P,b),t} = \frac{1}{\sqrt{m}}\, i^{wt(d_P)+2wt(b)}\, i^{tPt^\top+2bt^\top} \qquad (1)$$

where d_p denotes the main diagonal of P, and wt denotes the *Hamming weight* (the number of 1s in the binary vector). Note that all arithmetic in the expressions $tPt^\top + 2bt^\top$ and $wt(d_P)+2wt(b)$ takes place in the ring of integers modulo 4, since they appear only as exponents for i. Given P and b, the vector $tPt^\top + 2bt^\top$ is a codeword in the Delsarte-Goethals code. For a fixed matrix P, the 2^o columns $A_{(P,b)}$ ($b \in \mathbb{F}_2^o$) form an orthonormal basis Γ_P that can also be obtained by postmultiplying the Walsh-Hadamard basis by the unitary transformation $\operatorname{diag}\left[i^{tPt^\top}\right]$.

Throughout the rest of this section let $\mathbf{1}$ denote the all-one vector. Also let Φ denote the *unnormalized* DG frame, i.e., $A = \frac{1}{\sqrt{m}}\Phi$. We use the following lemmas to show that the Delsarte-Goethals frames are low-coherence tight-frames. First we prove that the columns of the r^{th} Delsarte-Goethals sensing matrix form a group under pointwise multiplication.



(a) horizontal (b) vertical (c) other

Fig. 1. Examples of images classified as horizontal (a), vertical (b), and other (c) using the measurement domain SVM classifier with a $DG(11,0)$ sensing matrix.

Lemma 4.1. *Let $\mathcal{G} = \mathcal{G}(o,r)$ be the set of unnormalized columns $\Phi_{(P,b)}$ where*

$$\phi_{(P,b),t} = i^{wt(d_P)+2wt(b)} i^{tPt^\top+2bt^\top}, \text{ where } t \in \mathbb{F}_2^o$$

where $b \in \mathbb{F}_2^o$ and where the binary symmetric matrix P varies over the Delsarte-Goethals set $DG(o,r)$. Then $\mathcal{G}$ is a group of order $2^{(r+2)o}$ under pointwise multiplication.

Proof. The proof of Lemma 4.1 is based on the construction of the DG frames, and is provided in [8]. □

Next we bound the worst-case coherence and average-coherence of the Delsarte-Goethals frames.

Corollary 4.1. *Let A be an $m \times n$ $DG(o,r)$ frame whose column entries are defined by* (1)*. Then $\mu \leq \frac{2^r}{\sqrt{m}}$, and $\nu = \frac{1}{n-1}$.*

Proof. The proof of Lemma 4.1 is based on the group properties of the DG frames, and is provided in [8]. □

Lemma 4.2. *Let A be a $DG(o,r)$ frame. Then A is a tight-frame with redundancy $\frac{n}{m}$.*

Proof. Let t and t' be two indices in $[m]$. We calculate the inner-product between the rows indexed by t and t'. It follows from Equation (1) that the inner-product can be written as

$$\sum_{P,b} a_{(P,b),t}\overline{a_{(P,b),t'}} = \frac{1}{m}\sum_{P,b} i^{tPt^\top - t'Pt'^\top + 2bt^\top - 2bt'^\top}$$
$$= \frac{1}{m}\left(\sum_P i^{tPt^\top - t'Pt'^\top}\right)\left(\sum_b (-1)^{b(t\oplus t')^\top}\right).$$

Therefore, since the columns of the matrix form a group under pointwise multiplication, if $t \neq t'$ then the inner-product is zero, and is $\frac{n}{m}$ otherwise. □

4.2.1. Compressed Learning via the Delsarte-Goethals Frames

Since Delsarte-Goethals frames have optimal worst-case and average-case coherence values, the measurement domain SVM classifier is near-optimal.

Theorem 4.1. *Let o be an odd integer, and let $r \leq \frac{o-1}{2}$. Let A be an $m \times n$ DG frame with $m = 2^o$, and $n = 2^{(r+2)o}$. Let $\boldsymbol{w_0}$ denote the data-domain oracle classifier. Also let $S \doteq \langle(\boldsymbol{x_1}, \mathrm{l}_1), \cdots, (\boldsymbol{x_M}, \mathrm{l}_M)\rangle$ represent M training examples in the data domain, and let $AS \doteq \langle(\boldsymbol{y_1}, \mathrm{l}_1), \cdots, (\boldsymbol{y_M}, \mathrm{l}_M)\rangle$ denote the representation of the training examples in the measurement domain. Finally let $\hat{\boldsymbol{z}}_{AS}$ denote the measurement domain SVM classifier trained on AS. Then there exist a universal constant C such that if*

$$m \geq \left(\frac{2^{r+1}C\log n}{\varepsilon_1}\right)^2, \quad \text{and} \quad k \leq \min\left\{\frac{m\varepsilon_1^2}{(2C)^2\log n}, n^{\frac{2}{3}}\right\}$$

then with probability at least $1 - \frac{6}{n}$, $H_{\mathcal{D}}(\hat{\boldsymbol{z}}_{AS}) - H_{\mathcal{D}}(\boldsymbol{w_0})$ is at most

$$O\left(R\|\boldsymbol{w_0}\|\sqrt{\left(\frac{2^r(\log M + \log n)}{\sqrt{m}} + \sigma + \frac{(1+\varepsilon_1)\log n}{M}\right)}\right). \quad (2)$$

Proof. The proof uses the probabilistic method, and the fact that the SVM decision is only based on the inner products between the test example and the support vectors. The proof shows that with overwhelming probability these inner products are approximately preserved by the projection matrix, and is provided in [9]. □

Remark 4.1. *Theorem 4.1 guarantees that larger measurement domain dimension m leads to lower measurement domain classification loss. In other words $H_{\mathcal{D}}(\hat{\boldsymbol{z}}_{AS})$ is bounded*

SVM	# of "horizontal"s	# of "vertical"s	# of "others"
Data Domain	14	18	23
Measurement Domain	12	15	28

Table 1. Comparison between the classification results of the SVM classifier in the data domain and the SVM classifier in the measurement domain.

by

$$H_{\mathcal{D}}(\boldsymbol{w_0})+\tilde{O}\left(\left(\frac{(\log M+\log n)^2}{m}\right)^{\frac{1}{4}}+\left(\frac{\log n}{M}\right)^{\frac{1}{2}}+\sigma^{\frac{1}{2}}\right).$$

Application: texture classification. Finally we demonstrate an application of compressed learning in texture classification. In texture classification, the goal is to classify images into one of the "horizontal", "vertical", or "other" classes. The information about the the direction of an image is stored in the horizontal and vertical wavelet coefficients of that image. Therefore, an SVM classifier in the data (pixel or wavelet) domain would provide high texture classification accuracy. Here we show that an SVM classifier trained directly over the compressively sampled images also has high performance.

We used the Brodatz texture database [10] which contains 111, 128×128 images. First we divided the dataset into 56 training images and 55 test images, and trained an SVM classifier from the 128×128 images. The images were then projected to a 2^{11} dimensional space using a $DG(11, 0)$ matrix. We then used the same procedure to train the measurement domain SVM classifier and classified the images accordingly. Table 1 compares the classification results of the SVM classifier in the data domain and the SVM classifier in the measurement domain. Figure 1 demonstrates examples of images in each class. The measurement domain classifier misclassifies 3 "horizontal" images, 3 "vertical" images and 1 "others" image. Therefore, the relative classification error of the measurement domain SVM classifier is $\frac{|14-11|+|18-15|+|23-22|}{55} \approx 12.7\%$.

5. CONCLUSION

In this paper we introduced compressed learning, a *linear* dimensionality reduction technique for measurement-domain pattern recognition in compressed sensing applications. We showed that a large family of compressed sensing matrices satisfy the required properties.

We again emphasize that the dimensionality reduction has been studied for a long time in many different communities. In particular, the development of theory and methods that cope with the curse of dimensionality has been the focus of the machine learning community for at least 20 years (e.g., SVM, complexity-regularization, model selection, boosting, aggregation, etc). The compressed learning approach of this section is most beneficial in compressed sensing applications. The reason is that compressed sensing already projects the data to some low dimensional space, and therefore the dimensionality reduction can be done as fast and efficiently as the state-of-the art sensing methods are.

References

[1] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 32(2), pp. 210-227*, 2009.

[2] E. Kushilevitz, Ostrovsky, and Y. Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing, Vol. 30(2), pp. 457-474*, 2000.

[3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Proceedings of the 28th annual ACM symposium on Theory of computing (STOC), pp. 20-29*, 1996.

[4] D. A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC), pp. 563-568*, 2008.

[5] J. Han, S. McKenna, and R. Wang. Regular texture analysis as statistical model selection. In *ECCV (4)*, volume 5305 of *Lecture Notes in Computer Science*, pages 242–255, 2008.

[6] C. J.C. Burgess. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[7] W. Bajwa, R. Calderbank, and S. Jafarpour. Model Selection: Two Fundamental Measures of Coherence and Their Algorithmic Significance. *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2010.

[8] R. Calderbank, S. Howard, and S. Jafarpour. Construction of a large class of Matrices satisfying a Statistical Isometry Propery. *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Compressive Sensing, Vol. 4(2), pp. 358-374*, 2010.

[9] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *Available at http://dsp.rice.edu/cs*, 2009.

[10] Brodatz Texture Database. *available at http://www.ux.uis.no/∼tranden/brodatz.html*.