Compressed Learning with Regular Concept*

Jiawei Lv[1], Jianwen Zhang[1], Fei Wang[2], Zheng Wang[1], and Changshui Zhang[1]

[1] State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
{lvjw05,jw-zhang06,w-z04}@mails.thu.edu.cn, zcs@mail.thu.edu.cn
[2] Department of Statistical Science, Cornell University
fw83@cornell.edu

Abstract. We revisit compressed learning in the PAC learning framework. Specifically, we derive error bounds for learning halfspace concepts with compressed data. We propose the *regularity* assumption over a pair of concept and data distribution to greatly generalize former assumptions. For a *regular* concept we define a *robust factor* to characterize the margin distribution and show that such a factor tightly controls the generalization error of a learned classifier. Moreover, we extend our analysis to the more general linearly non-separable case. Empirical results on both toy and real world data validate our analysis.

1 Introduction

The recent years have witnessed a surge of interest in *compressed learning* [2], i.e., learning with randomly projected data (compressed data) instead of original data. Compressed learning is necessary in two aspects: efficiency and privacy [13]. On one hand, learning with compressed data saves considerable running time and storage since random projection can effectively reduce the dimension of data. On the other hand, compressed learning can also be served as an important alternative for protecting data privacy. For example, in health care, security and finance related applications [9], data often contain sensitive information. Private database owners are only permitted to provide analyst with factitiously perturbed data rather than the original data. Random projection is a commonly used method to mask the original appearance of data for such privacy concerns [9]. In these scenarios, learning and analysis is only permitted to carried out on those randomly projected data. Existing works dealing with compressed data cover a variety of topics in machine learning such as classification, regression and manifold learning [1, 2, 4, 7, 10, 12, 13]. In this paper, we concentrate on the classification case.

A key issue in classification with compressed data is the learnability, whether the concept in the original space can be accurately learnt using the randomly projected data. There are two representative works on this problem.

* Supported by NSFC (Grant No. 60975003) and State Key Science and Technology Project on Marine Carbonate Reservoir Characterization (2008ZX05004-006).

M. Hutter et al. (Eds.): ALT 2010, LNAI 6331, pp. 163–178, 2010.

One is the loss function based analysis presented in [2]. They analyze the hinge loss of the linear classifier learned by support vector machine on compressed data and show the hinge loss of the compressedly learned classifier would not deviate much from that of the classifier learned on original data. However, they cannot guarantee the two classifiers have similar generalization error rates.

The other directly addresses the generalization error of a learned classifier on compressed data in the PAC framework [1]. The key factor affecting the generalization error of a learned classifier is the *margin distribution* of a concept, where the margin is defined as the distance between a sample and the separating boundary [6]. Since random projection perturbs data, a sample will very likely be wrongly classified if its margin is tiny. Therefore the learnability can only be guaranteed with some restrictions over the pair of concept and data distribution such that original data with small margins are of a nonessential proportion. Arriaga and Vempala [1] introduce the ℓ-robustness assumption ($\ell > 0$) which requires the margin of every sample is at least ℓ. Since random projection approximately preserves the margin of a finite set of samples [1], compressed learning of an ℓ-robust concept is in fact the classical problem of learning with a margin. They further show that an ℓ-robust halfspace concept can be accurately learned simply by a perceptron. However, the ℓ-robustness assumption is so restrictive that every halfspace concept is not ℓ-robust under many commonly adopted assumptions on data distributions (e.g., normal distributions and uniform distributions). Moreover, in real world problems, compressed data cannot always be well separated with a margin.

In this paper, we propose the *regularity* assumption to relax the *ℓ-robust* assumption to a more general case, under which the learnability of compressed learning halfspace concepts is revisited. The regularity assumption also imposes restrictions over a pair of concept and data distribution, but only requires that "almost" every sample has a nonzero margin. For example, a halfspace concept is regular with any continuous distribution. Hence we can work with more data distributions. However, under this relaxed assumption, compressed data may be wrongly classified since margins in the original space can be arbitrarily small. Therefore, we further define the *robust factor* for each regular concept to characterize the margin distribution. Under the regularity assumption, we revisit the problem of learning halfspace concepts with compressed data. We use the voted-perceptron algorithm proposed by Freund and Schapire [5] to perform the learning task on compressed data. Generalization error bounds based on the robust factor are derived for a learned classifier. We show that under some reasonable conditions on the robust factor, a regular halfspace concept can be accurately learned. The error analysis is also extended to the linearly non-separable case. Numerical experiments validate our analysis.

2 Preliminaries

The training set is denoted as $S = \{(x_1, y_1), ..., (x_m, y_m)\}$. $x_i \in \mathbb{R}^n$ is sampled independently from an unknown distribution $\mathcal{D}$ in $\mathbb{R}^n$ and then normalized onto the unit ball. The label $y_i \in \{-1, +1\}$ is given by an unknown halfspace concept $w \in \mathbb{R}^n$, i.e., $y_i = \operatorname{sign}\left(w^T x_i\right)$. Assume $\|w\| = 1$, where $\|\cdot\|$ is the Euclidean norm. R is a matrix of size $k \times n$. $w' = Rw$ is the projection of w under R. We use $w \cdot x$ to denote the inner product between w and x. $v(x) = |w \cdot x|$ is the distance between x and the hyperplane w. This is the margin of x with respect to w as used in [6]. The generalization error of a classifier $f(x)$ under distribution $\mathcal{D}$ is defined as

$$\mathrm{err}_{\mathcal{D}}(f) = \mathscr{P}_{\mathcal{D}}\{x : f(x) \neq y\},$$

where $\mathscr{P}_{\mathcal{D}}$ is the probability measure of $\mathcal{D}$.

2.1 Random Projection

Random projection is the technique of projecting a set of samples into a lower dimensional space by a random matrix. One of the most important properties of random projection is that it approximately preserve distance, which is stated by the following Johnson-Lindenstrauss lemma [1, 3, 8].

Theorem 1 (Johnson-Lindenstrauss lemma [1, 3, 8]). *Let $u, v \in \mathbb{R}^n$. R is a $k \times n$ random matrix with entries chosen independently from $N(0, 1/k)$. $u' = Ru$, $v' = Rv$. For any ϵ with $0 < \epsilon < 1$,*

$$\mathscr{P}\left\{(1-\epsilon)\|u-v\|^2 \leq \|u'-v'\|^2 \leq (1+\epsilon)\|u-v\|^2\right\} \geq 1 - e^{-\frac{\epsilon^2 k}{8}}.$$

The statement of Theorem 1 is due to Arriaga and Vempala [1].

In Theorem 1, the probability $\mathscr{P}$ is taken over the randomness of the matrix R. Specifically, if denoting $\mathbb{P}$ as the set of all the possible random matrices as in Theorem 1, then $\mathscr{P}$ is the probability measure over $\mathbb{P}$ defined by the construction of R. The corresponding distribution is denoted as $\mathrm{RP}(k, n)$.

Theorem 1 has a direct corollary as follows, which states that random projection also approximately preserves inner product.

Corollary 2. *Let $u, v \in \mathbb{R}^n$ with $\|u\|, \|v\| \leq 1$. Let $R \sim RP(k, n)$ and u', v' be the projections of u, v under R. Then for any $\epsilon > 0$,*

$$\mathscr{P}\{u \cdot v - \epsilon \leq u' \cdot v' \leq u \cdot v + \epsilon\} \geq 1 - 2e^{-\frac{\epsilon^2 k}{8}}.$$

Theorem 1 and Corallary 2 are crucial to our analysis.

2.2 ℓ-Robust Concepts

Our definition of a *regular* concept is motivated by the ℓ-robust concept, which is firstly introduced in [1].

Definition 3 (ℓ-robust halfspace concepts). *A half-space concept $w \in \mathbb{R}^n$ in conjunction with a distribution $\mathcal{D}$ in $\mathbb{R}^n$ is said to be ℓ-robust ($\ell > 0$), if*

$$\mathscr{P}_{\mathcal{D}}\{x : |w \cdot x| \leq \ell\} = 0. \tag{1}$$

Note that ℓ-robustness is an assumption placed over the couple $(w, \mathcal{D})$. It requires almost surely the margin of a sample is greater than ℓ. This margin measures how much one can alter a sample value without affecting its label. Therefore, the margin distribution characterizes the robustness of a concept to noise in sample values.

3 Regular Concepts and Robust Factors

In this section, we define regular halfspace concepts to generalize their ℓ-robust counterparts. For a regular concept, a corresponding robust factor is defined to characterize its margin distribution.

3.1 Regular Concepts

The formal definition of a regular concept is as follows.

Definition 4 (regular halfspace concepts). *Let $(w, \mathcal{D})$ be a halfspace concept and data distribution pair in $\mathbb{R}^n$. $(w, \mathcal{D})$ is called regular, if*

$$\mathscr{P}_{\mathcal{D}}\{x : |w \cdot x| = 0\} = 0. \tag{2}$$

When $(w, \mathcal{D})$ is regular, we call w a regular concept with respect to $\mathcal{D}$ or simply a regular concept when there is no confusion about $\mathcal{D}$. Clearly, an ℓ-robust concept is also a regular concept. A regular concept w requires points on the separating hyperplane forming a set of measure zero under $\mathcal{D}$. Therefore, almost surely every sample has a nonzero margin. Since under any continuous distribution in $\mathbb{R}^n$ the volume of an $n-1$ dimensional simplex is zero, every hyperplane is regular if the data distribution is continuous. Moreover, $(w, \mathcal{D})$ is not regular if there exists an $\epsilon > 0$, such that $\mathscr{P}_{\mathcal{D}}\{x : |w \cdot x| = 0\} = \epsilon$. Therefore, w is not regular when distributions are those placing nonzero measures on the separating hyperplane. This type of concept is unstable to noise in sample values, since any slight perturbation of a sample lying on the separating boundary will make it wrongly classified. We will not consider them here.

There exists an important property for a regular concept, which leads to the definition of the robust factor. The proof is a simple use of the monotone property of probability measures.

Proposition 5. *$(w, \mathcal{D})$ is regular if and only if for any $\epsilon > 0$, there exists an $\ell > 0$, such that*

$$\mathscr{P}_{\mathcal{D}}\{x : |w \cdot x| \leq \ell\} \leq \epsilon. \tag{3}$$

3.2 Robust Factors

The real number ℓ in (3) plays a similar role as that in an ℓ-robust concept, both characterizing the margin distribution of a concept. With a given $\epsilon > 0$, it is reasonable to believe that a pair $(w, \mathcal{D})$ with a larger ℓ corresponding to (3) is more robust to noise. This inspires us to introduce the following definition of the robust factor for a regular concept.

Definition 6 (robust factor). *The halfspace concept w is regular in conjunction with $\mathcal{D}$ in $\mathbb{R}^n$. Denote*

$$L_\epsilon = \{\ell \in [0,1] : \ell \textit{ satisfies } \mathscr{P}_{\mathcal{D}}\{|w\cdot x| \le \ell\} \le \epsilon\}$$

for a given $\epsilon > 0$. Define $\phi(\epsilon) = \sup_{\ell \in L_\epsilon} \ell$ as the robust factor with respect to $(w, \mathcal{D})$.

It can be easily shown that $\phi(\epsilon) \in L_\epsilon$. When the distribution is continuous, $\phi(\epsilon)$ is simply the largest real number ℓ such that $\mathscr{P}_{\mathcal{D}}\{|w\cdot x| \le \ell\} = \epsilon$. See the following examples, where $\phi(\epsilon)$ can be explicitly expressed or bounded.

Example 7 (robust factor of uniform distribution on the unit sphere in $\mathbb{R}^3$). Let $w = [1,0,0]^T$ be the halfspace concept. $\mathcal{D}$ is the uniform distribution on the unit sphere in $\mathbb{R}^3$. Then $\phi(\epsilon) = \frac{\epsilon}{4\pi}$.

Generally, the robust factor of uniform distribution on the unit sphere in $\mathbb{R}^n$ has the following upper bounds.

Example 8. $w = [1,0,\cdots,0]^T \in \mathbb{R}^n\,(n > 3)$ is the halfspace concept. $\mathcal{D}$ is the uniform distribution on the unit sphere in $\mathbb{R}^n$. Then for any $\epsilon \in (0,1)$, we have

$$\phi(\epsilon) \le \begin{cases} \frac{\Gamma(\frac{n}{2})\epsilon}{8\pi^{(n-1)/2}} & n \text{ is even,} \\ \sin\left(\frac{\Gamma(\frac{n}{2})\epsilon}{4\pi^{(n-1)/2}}\right) & n \text{ is odd.} \end{cases}$$

Proof. Transforming to the spherical coordinate system, we can represent the n Cartesian coordinates as follows

$$\begin{aligned} x_1 &= \cos\theta_1, \\ x_2 &= \sin\theta_1 \cos\theta_2, \\ &\cdots, \\ x_{n-1} &= \sin\theta_1 \cdots \sin\theta_{n-2}\cos\theta_{n-1}, \\ x_n &= \sin\theta_1 \cdots \sin\theta_{n-2}\sin\theta_{n-1}, \end{aligned}$$

where $\theta_i \in [0,\pi]$ for $1 \le i \le n-2$ and $\theta_{n-1} \in [0, 2\pi]$. Furthermore, we have $|w\cdot x| = |\cos\theta_1|$. Let $\theta \in [0, \pi/2]$ such that $\cos\theta = \phi(\epsilon)$. Denote $A_{\phi(\epsilon)} = \mathscr{P}_{\mathcal{D}}\{|w\cdot x| \le \phi(\epsilon)\}$. We have

$$\begin{aligned} A_{\phi(\epsilon)} =& \int_{\theta_1=\theta}^{\pi-\theta}\int_{\theta_2=0}^{\pi}\cdots\int_{\theta_{n-2}=0}^{\pi}\int_{\theta_{n-1}=0}^{2\pi} dS \\ =& \int_{\theta}^{\pi-\theta} d\theta_1 \sin^{n-2}\theta_1 \int_0^\pi d\theta_2 \sin^{n-3}\theta_2 \cdots \int_0^\pi d\theta_{n-2}\sin\theta_{n-2}\int_0^{2\pi} d\theta_{n-1} \\ =& \frac{2\pi^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)}\int_\theta^{\pi-\theta} d\theta_1 \sin^{n-2}\theta_1, \end{aligned}$$

where dS is the area element of the n-sphere, i.e.,

$$dS = \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}\, d\theta_1 \cdots d\theta_{n-1},$$

and we obtain the last equality by using the surface area formula of the n-sphere. By the technique of integration by parts, we have the following recurrence formula:

$$\int_{\theta}^{\pi-\theta} \sin^{n-2} t\, dt = \frac{2\sin^{n-3}\theta\cos\theta}{n-2} + \frac{n-3}{n-2}\int_{\theta}^{\pi-\theta} \sin^{n-4} t\, dt.$$

Since $\theta \in [0, \pi/2]$,

$$\int_{\theta}^{\pi-\theta} \sin^{n-2} t\, dt \geq \frac{n-3}{n-2}\int_{\theta}^{\pi-\theta} \sin^{n-4} t\, dt. \tag{4}$$

If n is even, by (4), we have

$$\int_{\theta}^{\pi-\theta} \sin^{n-2} t\, dt \geq \frac{(n-3)(n-5)\cdots 1}{(n-2)(n-4)\cdots 2}\int_{\theta}^{\pi-\theta} dt = 2\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\left(\frac{\pi}{2}-\theta\right).$$

Therefore,

$$\epsilon = A_{\phi(\epsilon)} \geq \frac{4\pi^{(n-1)/2}}{\Gamma(\frac{n}{2})}\left(\frac{\pi}{2}-\theta\right).$$

We have $\frac{\pi}{2} - \theta \leq \frac{\Gamma(\frac{n}{2})\epsilon}{4\pi^{(n-1)/2}}$. And the robust factor can be bounded by

$$\phi(\epsilon) = \cos\theta \leq \sin\left(\frac{\Gamma(\frac{n}{2})\epsilon}{4\pi^{(n-1)/2}}\right).$$

If n is odd, we have

$$\int_{\theta}^{\pi-\theta} \sin^{n-2} t\, dt \geq \frac{(n-3)(n-5)\cdots 2}{(n-2)(n-4)\cdots 3}\int_{\theta}^{\pi-\theta} \sin t\, dt = 4\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\cos\theta.$$

Therefore,

$$\epsilon = A_{\phi(\epsilon)} \geq \frac{8\pi^{(n-1)/2}}{\Gamma(\frac{n}{2})}\cos\theta.$$

And this directly gives the upper bound of the robust factor, i.e.,

$$\phi(\epsilon) \leq \frac{\Gamma(\frac{n}{2})\epsilon}{8\pi^{(n-1)/2}}.$$

Thus the result follows.

The robust factor of an ℓ-robust concept satisfies $\phi(0) = \ell > 0$. This is a very strong condition and does not generally hold for a regular concept. We only assume $\phi(0) = 0$ in the following discussion. Hence, as $\epsilon \to 0$, $\phi(\epsilon)$ also decreases to 0. The speed that $\phi(\epsilon)$ approaches zero is an important characteristic of the margin distribution which greatly affects the learning result. We prefer those decreasing slowly near $\epsilon = 0$. This is because the robust factor with a smaller decreasing rate will eventually take a larger value when ϵ is sufficiently close to zero. Only for those concepts with their robust factors not decreasing too fast near $\epsilon = 0$, can we learn a classifier accurate enough. Based on this idea, we introduce the following additional assumption to control the decreasing rate of a robust factor. This assumption can be viewed as a slight variant of Tsybakov's noise condition [11].

Assumption 9. *Suppose there exist constants $0 < \epsilon_0 < 1$, $C > 0$ and $0 \leq \alpha < 1/4$ such that for every $\epsilon \in [0, \epsilon_0)$, $\phi(\epsilon)$ satisfies*

$$\phi(\epsilon) \geq C\epsilon^{\alpha}. \tag{5}$$

If the assumption holds, there does not exist such a neighborhood of the origin that $\phi(\epsilon) \leq \bar{C}\epsilon^{\beta}$ holds for all ϵ in the neighborhood, where $\bar{C} > 0$ and $\beta > \alpha$. And the rate of $\phi(\epsilon)$ decreasing to zero is no faster than that of a polynomial with order less than 1/4. Not all robust factors satisfy this assumption, e.g., the above two examples. We will give experimental results on toy data to show that failing to satisfy it will greatly affect the accuracy of the compressedly learned classifier. It will also be shown in the following section that if the assumption holds, an sufficiently accurate classifier can be learned based on compressed data.

4 Learning Regular Halfspace Concepts

In this section, we present error bounds for learning regular halfspace concepts with compressed data. We first summarize our main results then give the proofs.

4.1 Main Results

We use the voted-perceptron algorithm proposed in [5] as the base algorithm to learn from compressed data. The outputs of the algorithm is a weighted ensemble of linear classifiers. There is a parameter T in the algorithm which represents the number of iterations. In the following analysis, we simply set $T = 1$ for convenience. Note that our main results are not particular for the base algorithm, since the analysis does not depend on the detailed structure of the algorithm. In fact, any algorithm with a comparable generalization guarantee can be used to obtain similar results.

Theorem 10. *Let $w \in \mathbb{R}^n$ be a halfspace concept with $\|w\| = 1$. $\mathcal{D}$ is a distribution over $\mathbb{R}^n$. Suppose $(w, \mathcal{D})$ is regular. Let $R \sim RP(k, n)$ be a random matrix. S is the training set of size m. S' is its projection under R. For any given $\epsilon > 0$ and $\delta > 4\epsilon$, with probability at least $1-\delta$, the generalization error of the classifier $f(x)$ output by the voted-perceptron algorithm based on S' satisfies*

$$err_{\mathcal{D}}(f) \leq \frac{1}{\phi^2}\left(\frac{16(1+\epsilon)^2 + 16\delta}{m+1} + 40\epsilon + 18\delta\right), \tag{6}$$

if $k = \max\left\{\frac{32}{\phi^2}\log\frac{2}{\epsilon(\delta/4-\epsilon)}, \frac{8}{\epsilon^2}\log\frac{2(m+2)}{\delta}\right\}$ *where* $\phi = \phi\left(\epsilon^2\right)$*, provided* $m+1 \geq \frac{1}{2\epsilon^2}\log(8/\delta)$.

The generalization error depends critically on the robust factor. If the robust factor takes a larger value, the learned classifier will be more accurate. This is consistent with our intuition that, if samples own larger margins, the learning problem should be easier.

Also, as the sample size m approaches infinity, the upper bound approaches $\frac{c(\epsilon+\delta)}{\phi^2}$, where c is some constant. And the convergence rate is $1/m$. Hence, when training sample is enough, the accuracy of the compressedly learned classifier is determined by the intrinsic hardness of the problem, i.e., the margin distribution.

As a simple corollary of Theorem 10, it can be shown that under Assumption 9, the learned classifier $f(x)$ can be arbitrarily accurate by suitably choosing the dimension k and sample size m. This guarantees we can learn a regular halfspace concept with satisfactory accuracy on compressed data.

Corollary 11. *We invoke Assumption 9. Let* α *and* ϵ_0 *be the constants from Assumption 9. For any* $\epsilon > 0$*, let* $\epsilon_1 = \min\left\{\epsilon_0, c_1\left(\epsilon/2\right)^{\frac{1}{1-4\alpha}}\right\}$*, where* c_1 *is some fixed constant. If* δ *satisfies* $4\epsilon_1 < \delta < c_2\epsilon_1$ *for some constant* c_2*, then with probability at least* $1-\delta$*, the learned classifier* $f(x)$ *in Theorem 10 satisfies*

$$err_{\mathcal{D}}\left(f\right) \leq \epsilon,$$

if $k = \max\left\{\frac{32}{\phi^2}\log\frac{2}{\epsilon_1(\delta/4-\epsilon_1)}, \frac{8}{\epsilon_1^2}\log\frac{2(m+2)}{\delta}\right\}$ *and* $m \geq \max\left\{160/\phi^2, \frac{\log(8/\delta)}{2\epsilon_1^2}\right\}$*, where* $\phi = \phi\left(\epsilon_1^2\right)$.

Specifically, when $(w,\mathcal{D})$ is ℓ-robust, the sample complexity in Corollary 11 is $\tilde{O}\left(1/\ell^2\right)$ in terms of ℓ, the same order as Arriaga and Vempala's [1], where the $\tilde{O}(\cdot)$ notation hides the logarithmic factors. However our lower bound on the projection dimension k is $k \geq \Omega\left(\log m\right)$, which improves the bound $k \geq \Omega\left(m\right)$ obtained in [1] .

We will give the detailed proof of Theorem 10 in the following part of this section. The proof of Corollary 11 is direct and mainly technical. We omit it here.

4.2 Proofs

For a given w, define the product set $E \subset \mathbb{P} \times \mathbb{R}^n$ as

$$E = \left\{(R,x) : \text{sign}\left(w \cdot x\right) \neq \text{sign}\left(w' \cdot x'\right)\right\},$$

where $w' = Rw$, $x' = Rx$. For a point $x \in \mathbb{R}^n$, the x *cross section* of E is defined as $E_x = \left\{R : (R,x) \in E\right\} = \left\{R : \text{sign}\left(w \cdot x\right) \neq \text{sign}\left(w' \cdot x'\right)\right\}$. Similarly, the R cross section of E is $E_R = \left\{x : (R,x) \in E\right\} = \left\{x : \text{sign}\left(w \cdot x\right) \neq \text{sign}\left(w' \cdot x'\right)\right\}$. We first need to build some propositions and lemmas.

Lemma 12. *Given $w, x \in \mathbb{R}^n$ with $\|w\| = \|x\| = 1$, for any $\ell \in (0,1)$,*

$$\mathscr{P}\{E_x\} \leq 2e^{-\frac{\ell^2 k}{8}} + I(|w \cdot x| \leq \ell)$$

holds, where $I(\cdot)$ is the indicator function.

Proof. Decompose E_x according to whether $F = \{|w \cdot x| \leq \ell\}$ happens or not. Since $E_x \cap F \subseteq F$ and $E_x \cap F^c \subseteq \{|w \cdot x - w' \cdot x'| > \ell\}$, the lemma follows directly by Proposition 5 and Corollary 2.

Proposition 13. *Let $w \in \mathbb{R}^n$ be a halfspace concept. $\|w\| = 1$. $\mathcal{D}$ is a distribution over $\mathbb{R}^n$. Suppose $(w, \mathcal{D})$ is regular. $R \sim RP(k,n)$ and $w' \in \mathbb{R}^k$ is the projection of w under R. Then for any $0 < \epsilon < \delta$, we have*

$$\mathscr{P}\{err_{\mathcal{D}}(w') > \epsilon\} \leq \delta,$$

if $k \geq \frac{8}{\phi(\epsilon^2)^2} \log \frac{2}{\epsilon(\delta-\epsilon)}$.

Proof. Clearly, $|\mathrm{err}_{\mathcal{D}}(w) - \mathrm{err}_{\mathcal{D}}(w')| = \mathrm{err}_{\mathcal{D}}(w')$. We will first show that for a fixed $R \in \mathbb{P}$,

$$|\mathrm{err}_{\mathcal{D}}(w) - \mathrm{err}_{\mathcal{D}}(w')| \leq \mathscr{P}_{\mathcal{D}}\{E_R\}. \tag{7}$$

In fact by rewriting probability into an integral of the corresponding indicator function, we have

$$\begin{aligned} \left|\mathrm{err}_{\mathcal{D}}(w) - \mathrm{err}_{\mathcal{D}}\left(w'\right)\right| &= \left|\int I(\mathrm{sign}(w \cdot x) \neq y)\ d\mathscr{P}_{\mathcal{D}} - \int I\left(\mathrm{sign}\left(w' \cdot x'\right) \neq y\right)\ d\mathscr{P}_{\mathcal{D}}\right| \\ &\leq \int \left|I(\mathrm{sign}(w \cdot x) \neq y) - I\left(\mathrm{sign}\left(w' \cdot x'\right) \neq y\right)\right|\ d\mathscr{P}_{\mathcal{D}} \\ &= \int I(E_R)\ d\mathscr{P}_{\mathcal{D}} = \mathscr{P}_{\mathcal{D}}\{E_R\}. \end{aligned}$$

The expectation of $\mathrm{err}_{\mathcal{D}}(w')$ with respect to the randomness of the random matrix is

$$\begin{aligned} \mathscr{E}\{\mathrm{err}_{\mathcal{D}}(w')\} &= \int \mathrm{err}_{\mathcal{D}}(w')\ d\mathscr{P} = \int |\mathrm{err}_{\mathcal{D}}(w) - \mathrm{err}_{\mathcal{D}}(w')|\ d\mathscr{P} \\ &\leq \int \mathscr{P}_{\mathcal{D}}\{E_R\}\ d\mathscr{P} = \int\int I(E)\ d\mathscr{P}_{\mathcal{D}}\ d\mathscr{P} \\ &= \int\int I(E)\ d\mathscr{P}\ d\mathscr{P}_{\mathcal{D}} = \int \mathscr{P}\{E_x\}\ d\mathscr{P}_{\mathcal{D}} \\ &\leq \int 2e^{-\frac{\phi\left(\epsilon^2\right)^2 k}{8}} + I\left(|w \cdot x| \leq \phi\left(\epsilon^2\right)\right)\ d\mathscr{P}_{\mathcal{D}} \\ &\leq\ 2e^{-\frac{\phi\left(\epsilon^2\right)^2 k}{8}} + \epsilon^2, \end{aligned}$$

where the first equality of the third line is obtained by Fubini's Theorem and the last inequality but one is because of Lemma 12. Therefore, by Markov's Inequality, we have

$$\mathscr{P}\left\{\mathrm{err}_{\mathcal{D}}\left(w'\right) > \epsilon\right\} \leq \frac{\mathscr{E}\{\mathrm{err}_{\mathcal{D}}\left(w'\right)\}}{\epsilon} \leq \frac{2}{\epsilon} e^{-\frac{\phi\left(\epsilon^2\right)^2 k}{8}} + \epsilon.$$

By solving $\frac{2}{\epsilon} e^{-\frac{\phi\left(\epsilon^2\right)^2 k}{8}} + \epsilon \leq \delta$ for k, the proposition follows.

We also have the following result. The proof is very similar with that of Proposition 13. The key idea is using Fubini's Theorem to change the integration orders. We omit the details here.

Proposition 14. *Let $w \in \mathbb{R}^n$ be a halfspace concept. $\|w\| = 1$. $\mathcal{D}$ is a distribution over $\mathbb{R}^n$. Suppose $(w, \mathcal{D})$ is regular. x is a random sample with distribution $\mathcal{D}$. $R \sim RP(k, n)$. Let $w', x' \in \mathbb{R}^k$ be the projection of w, x under R, respectively. Then for any $0 < \epsilon < 1/2, \delta > \epsilon$, if $k \geq \frac{32}{\phi^2} \log \frac{2}{\epsilon(\delta - \epsilon)}$,*

$$\mathscr{P}\left\{\mathscr{P}_{\mathcal{D}}\left\{\left|w' \cdot x'\right| \leq \phi/2\right\} > \epsilon\right\} < \delta \tag{8}$$

holds where $\phi = \phi\left(\epsilon^2\right)$.

We also need the following error bound of the base learning algorithm due to Freund and Schapire [5].

Theorem 15. *Let S be a set of m samples with $\|x_i\| \leq r$. Let (x_{m+1}, y_{m+1}) be a test sample. For $h \in \mathbb{R}^n$, $\|h\| = 1$ and $\gamma > 0$, let*

$$D_{h,\gamma} = \sqrt{\sum_{i=1}^{m+1} \xi_i^2} = \sqrt{\sum_{i=1}^{m+1} \max\left\{0, \gamma - y_i\left(h \cdot x_i\right)\right\}^2}.$$

Then the probability (over the choice of all $m + 1$ samples) that the voted-perceptron algorithm with $T = 1$ does not predict y_{m+1} on the test sample x_{m+1} is at most

$$\mathscr{E}_{\mathcal{D}}\left\{\frac{2}{m+1} \inf_{\|h\|=1; \gamma>0} \left(\frac{r + D_{h,\gamma}}{\gamma}\right)^2\right\},$$

where $\mathscr{E}_{\mathcal{D}}\{\cdot\}$ is the expectation over the $m + 1$ samples.

Now we are in the position of presenting the complete proof of Theorem 10.

Proof of Theorem 10. Let (x'_{m+1}, y_{m+1}) be the projection of the test sample (x_{m+1}, y_{m+1}). Take $k = \max\left\{\frac{32}{\phi^2} \log \frac{2}{\epsilon(\delta/4 - \epsilon)}, \frac{8}{\epsilon^2} \log \frac{2(m+2)}{\delta}\right\}$, where $\phi = \phi\left(\epsilon^2\right)$. By Proposition 13 and Proposition 14,

$$\mathscr{P}\left\{\mathrm{err}_{\mathcal{D}}\left(w'\right) \geq \epsilon\right\} \leq \delta/4,$$
$$\mathscr{P}\left\{\mathscr{P}_{\mathcal{D}}\left\{\left|w' \cdot x'\right| \geq \phi/2\right\} \geq \epsilon\right\} \leq \delta/4$$

both holds. Moreover, by Theorem 1, $\|w\| = 1$, and $\|x_i\| = 1$, $1 \leq i \leq m+1$, we have $\mathscr{P}\{\|w'\| \geq 1+\epsilon\} \leq \frac{\delta}{2(m+2)}$ and $\mathscr{P}\{\|x_i'\| \geq 1+\epsilon\} \leq \frac{\delta}{2(m+2)}$, for $1 \leq i \leq m+1$. Define the "good" set of random matrix as:

$$\begin{aligned} G = &\{R : \mathrm{err}_{\mathcal{D}}(w') \leq \epsilon\} \cap \{R : \mathscr{P}_{\mathcal{D}}\{|w' \cdot x'| \leq \phi/2\} \leq \epsilon\} \\ &\cap \{R : \|w'\| \leq 1+\epsilon\} \cap \left\{R : \max_{1\leq i\leq m+1} \|x_i'\| \leq 1+\epsilon\right\}. \end{aligned}$$

Hence $\mathscr{P}\{G\} \geq 1-\delta$. Fix a $R \in G$ in the following. S' is the projection of S under R. Denote $A' = S' \cup \left\{(x'_{m+1}, y_{m+1})\right\}$. The empirical error of w' on A' is

$$\mathrm{err}_{A'}(w') = \frac{1}{m+1} \sum_{i=1}^{m+1} I\left(\mathrm{sign}\left(w' \cdot x_i'\right) \neq y_i\right).$$

By the Chernoff bound, we have

$$\mathscr{P}_{\mathcal{D}}\{|\mathrm{err}_{A'}(w') - \mathrm{err}_{\mathcal{D}}(w')| \geq \epsilon\} \leq 2e^{-2(m+1)\epsilon^2} < \frac{\delta}{4}.$$

Since $R \in G$, $\mathrm{err}_{\mathcal{D}}(w') \leq \epsilon$. Therefore, we obtain

$$\mathscr{P}_{\mathcal{D}}\{(m+1)\mathrm{err}_{A'}(w') \leq 2\epsilon(m+1)\} \geq 1 - \frac{\delta}{4},$$

provided $m+1 \geq \frac{\log(8/\delta)}{2\epsilon^2}$. Similarly, we can also bound the number of samples correctly classified by w' but with margin less than $\phi/2$, i.e.,

$$\begin{aligned} &\mathscr{P}_{\mathcal{D}}\left\{\sum_{i=1}^{m+1} I\left(0 < y_i\left(w' \cdot x_i'\right) \leq \phi/2\right) \leq 2\epsilon(m+1)\right\} \\ \geq &\mathscr{P}_{\mathcal{D}}\left\{\sum_{i=1}^{m+1} I\left(|w' \cdot x_i'| \leq \phi/2\right) \leq 2\epsilon(m+1)\right\} \geq 1 - \frac{\delta}{4}. \end{aligned}$$

Setting $\gamma = \phi/2$, $D_{h,\gamma}^2$ can be upper bounded by $(m+1)(1+\epsilon+\phi/2)^2$. Hence

$$\inf_{\|h\|=1,\gamma} \left(\frac{r + D_{h,\gamma}}{\gamma}\right)^2 \leq \frac{1+\epsilon+2(m+1)\left(1+\epsilon+\phi/2\right)^2}{\phi^2/4},$$

if $m \geq 4$. What's more, D has a tighter bound. With probability at least $1-\delta/2$ over the randomness of all $m+1$ samples, we have

$$\begin{aligned} D_{w',\phi/2}^2 = &\sum_{i=1}^{m+1} \max\left\{0, \phi/2 - y_i\left(w' \cdot x_i'\right)\right\}^2 \\ \leq &(1+2\epsilon+\epsilon^2+\phi/2)^2 \sum_{i=1}^{m+1} I\left(\mathrm{sign}\left(w' \cdot x_i'\right) \neq y_i\right) \\ &+ \phi^2/4 \sum_{i=1}^{m+1} I\left(0 < y_i\left(w' \cdot x_i'\right) \leq \phi/2\right) \\ \leq &2\epsilon(m+1)\left(4 + 2\phi + \phi^2/2\right). \end{aligned}$$

Note that $D_{w'/\|w'\|,\phi/(2\|w'\|)} = D_{w',\phi/2}/\left\|w'\right\|$. Therefore, if $m \geq 2/\epsilon$, with probability at least $1-\delta/2$,

$$\inf_{\|h\|=1,\gamma}\left(\frac{r+D_{h,\gamma}}{\gamma}\right)^2 \leq \left(\frac{(1+\epsilon)^2+D_{w',\phi/2}}{\phi/2}\right)^2 \leq \frac{(1+\epsilon)^4+2D^2_{w',\phi/2}}{\phi^2/4}$$
$$\leq \frac{(1+\epsilon)^4+4\epsilon(m+1)(4+2\phi+\phi^2/2)}{\phi^2/4},$$

Hence, we can finally bound the error rate

$$\mathrm{err}_{\mathcal{D}}(f) \leq \mathscr{E}_{\mathcal{D}}\left\{\frac{2}{m+1}\inf_{\|h\|=1;\gamma>0}\left(\frac{r+D_{h,\gamma}}{\gamma}\right)^2\right\}$$
$$\leq \left(1-\frac{\delta}{2}\right)\frac{(1+\epsilon)^4+4\epsilon(m+1)(4+2\phi+\phi^2/2)}{(m+1)\phi^2/8}$$
$$+\frac{\delta}{2}\frac{1+\epsilon+2(m+1)\left(1+\epsilon+\phi/2\right)^2}{(m+1)\phi^2/8}$$
$$\leq \frac{16}{m+1}\frac{(1+\epsilon)^2+\delta}{\phi^2}+\frac{40\epsilon+18\delta}{\phi^2}.$$

5 Linearly Non-separable Case

Our analysis can be further extended to the case when data are linearly non-separable in the original space. We would like the compressedly learned classifier to be not much worse than the best linear classifier in the original space. We need to generalize the definition of regularity to a general linear classifier. Here, we only consider linear classifiers passing through the origin.

Definition 16. *A linear classifier h with a distribution $\mathcal{D}$ in $\mathbb{R}^n$ is regular, if*

$$\mathscr{P}_{\mathcal{D}}\{x : |h \cdot x| = 0\} = 0. \tag{9}$$

The definition is the same as that of a regular concept. However, here $\mathrm{err}_{\mathcal{D}}(h)$ is generally nonzero. Let $\hat{h} = \arg\min_{h\in\mathbb{R}^n}\mathrm{err}_{\mathcal{D}}(h)$ with η the minimal error rate. The following result bounds the generalization error of the classifier $f(x)$ learned on compressed data in terms of η.

Theorem 17. *$\mathcal{D}$ is a distribution in $\mathbb{R}^n$. Let $\hat{h} \in \mathbb{R}^n$ be the linear classifier with the minimal error rate η under $\mathcal{D}$. $\left\|\hat{h}\right\| = 1$. Suppose $(\hat{h},\mathcal{D})$ is regular. $R \sim RP(k,n)$. S is the training set of size m. Let S' be the projection of S under R. For any given $\epsilon > 0$ and $\delta > 8\epsilon$, with probability at least $1-\delta$, the generalization error of the classifier $f(x)$ output by the voted-perceptron algorithm based on S' satisfies,*

$$err_{\mathcal{D}}(f) \leq \frac{1}{\phi^2}\left(\frac{4(1+\epsilon)^2+2\delta}{m+1}+10(\epsilon+\eta)+8\delta\right),$$

if $k = \max\left\{\frac{8}{\phi(\epsilon^2)^2}\log\frac{2}{\epsilon(\delta/8-\epsilon)}, \frac{32}{\phi^2}\log\frac{2}{\epsilon(\delta/8-\epsilon)}, \quad \frac{8}{\epsilon^2}\log\frac{2(m+1)}{\delta}\right\}$,
where $\phi = \phi\left((\epsilon+\eta)^2\right)$.

This result shows that if the margin distribution places little mass on small margins, one can learn a linear classifier with compressed data which would not be much too worse than the best linear classifier one can learned with original data. When m is sufficiently large, the upper bound approximates $\frac{c(\epsilon+\eta+\delta)}{\phi^2}$, and the convergence rate is $1/m$.

The proof is similar with that of Theorem 10. We omit it here and provide it in the supplementary material.

6 Experiments

In this section, we provide experimental results on synthesize data. The purpose of these experiments is to test how different distributions affect the accuracy of a classifier trained on compressed data and whether these effects are consistent with our analysis. We first introduce the experimental setups, and then provide the detailed experimental results.

6.1 Experimental Setup

For each data set used in our experiment, we randomly choose 90% of the samples as the training set, and the rest is used as the testing set. We then train two classifiers using the base algorithm on the original training set and compressed training set respectively, where the compressed training set is obtained by projecting the original training set to a k-dimensional space with a random matrix. Finally the classification accuracy is evaluated on the original and compressed testing set with the originally and compressedly trained classifier. To reduce the training bias, we repeat the above procedure for N rounds and report the averaged results. In this way, we totally trained $M \times N$ classifiers on compressed data, and the averaged accuracy of compressedly trained classifiers are also reported to compare with that of the classifiers trained on original data. We choose several different values of dimension k and plot the test accuracy curve as a function of the dimension. In the experiments, we set $M = 50$, $N = 5$ and $T = 10$ in the base algorithm.

6.2 Numerical Experiments

We synthesize data to test two aspects of our analysis: the effect of failing to satisfy Assumption 9 and the effect of different values of robust factors, on the compressedly learned classifiers.

First, we show that when Assumption 9 does not hold, the compressedly learned classifier could be significantly worse than the classifier learned on the original data. From Example 8, it can be seen that the robust factor of uniform distribution on the unit sphere in $\mathbb{R}^n$ fails to satisfy the assumption. We test this type of

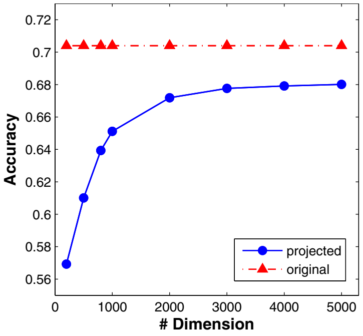

Fig. 1. Classification accuracy result on the uniform distribution over the unit n-sphere

distribution with $n = 1000$. The toy data set consists of 1000 i.i.d. samples generated from the distribution. $w = [1, 0, \cdots, 0]^T$ is the halfspace concept to learn. The averaged classification accuracy result is shown in Figure 1, which shows that the compressedly learned classifier is much worse than the classifier trained on original data, with a non-negligible error gap between the two classifiers even when the dimension k is greatly larger than the original dimension n.

Second, we test the effect of the robust factor on the generalization error of a learned classifier. Specifically, we want to show that if the robust factor of a distribution takes a larger value, the learned classifier will be more accurate. We choose $w = [1, 0, \cdots, 0]^T$ as the common halfspace concept and obtain a sequence of distributions with different robust factors near $\epsilon = 0$ by modifying a 1000-dimensional uniform distribution on the unit sphere as follows. When a sample x is generated from the uniform distribution, we compute the value $v(x) = |w \cdot x|$. If $v(x)$ is less than a threshold (0.01 in our experiments) we reject this sample with probability $1 - p$. Otherwise we simply accept it. The robust factor of the obtained distribution has a larger value near $\epsilon = 0$ than that of the

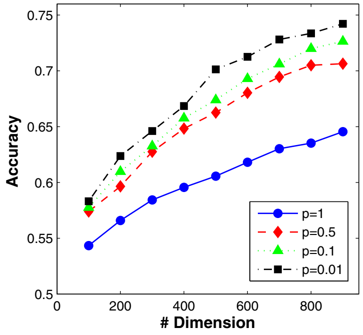

Fig. 2. Accuracy curves for four distributions with different robust factors. Small values of p give relatively large values of robust factor near $\epsilon = 0$.

uniform distribution. A smaller p gives a larger value of the robust factor near $\epsilon = 0$. We test four different values of p: 1, 0.5, 0.1, 0.01. From each distribution, we generate 1000 samples. The averaged classification accuracy result is shown in Figure 2. For a common k, a distribution with a larger robust factor near the origin results in a more accurate classifier. This results approve our analysis using robust factor to bound the error rate of a learned classifier.

7 Conclusion

In this paper, we study the problem of learning regular halfspace concepts with compressed data. We notice that the hardness of compressed learning is captured by the margin distribution. Therefore, we define the robust factor to characterize the margin distribution of a regular concept. We show that the generalization error of a compressedly learned classifier is tightly bounded in terms of the robust factor. Our analysis is also extended to the linearly non-separable case. Both theoretical and experimental results are provided to show that under certain conditions, learning halfspace concepts accurately with compressed data is possible.

References

[1] Arriaga, R.I., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projection. In: FOCS 1999: Proc. of the 40th Foundations of Computer Science (1999)

[2] Calderbank, R., Jafarpour, S., Schapire, R.: Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Princeton University (2009)

[3] Dasgupta, S., Gupta, A.: An elementary proof of a theorem of johnson and lindenstrauss. Random Structures and Algorithms 22(1), 60–65 (2003)

[4] Freund, Y., Dasgupta, S., Kabra, M., Verma, N.: Learning the structure of manifolds using random projections. In: Advances in Neural Information Processing Systems, vol. 20, pp. 473–480. MIT Press, Cambridge (2008)

[5] Freund, Y., Schapire, R.E.: Large margin classification using the perceptron algorithm. Mach. Learn. 37(3), 277–296 (1999)

[6] Garg, A., Har-Peled, S., Roth, D.: On generalization bounds, projection profile, and margin distribution. In: ICML 2002: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 171–178 (2002)

[7] Hegde, C., Wakin, M., Baraniuk, R.: Random projections for manifold learning. In: Advances in Neural Information Processing Systems, vol. 20, pp. 641–648. MIT Press, Cambridge (2008)

[8] Johnson, W., Lindenstrauss, J.: Extensions of lipschitz maps into a hilbert space. Contemp. Math. 26, 189–206 (1984)

[9] Liu, K., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans. on Knowl. and Data Eng. 18(1), 92–106 (2006); Senior Member-Kargupta, Hillol

[10] Maillard, O., Munos, R.: Compressed least-squares regression. In: Advances in Neural Information Processing Systems, vol. 21. MIT Press, Cambridge (2009)

[11] Tsybakov, A.B.: Optimal aggregation of classifiers in statistical learning. Ann. Statist. 32, 135–166 (2004)

[12] Wang, F., Li, P.: Efficient non-negative matrix factorization with random projections. In: The 10th SIAM International Conference on Data Mining, pp. 281–292 (2010)

[13] Zhou, S., Lafferty, J., Wasserman, L.: Compressed regression. In: Advances in Neural Information Processing Systems, vol. 20, pp. 1713–1720. MIT Press, Cambridge (2008)