Randomness and Computation
Joint Workshop "New Horizons in Computing" and "Statistical Mechanical Approach to Probabilistic Information Processing" (18-21 July, 2005, Sendai, Japan)

Random Projection and Its Application to Learning

Tatsuya Watanabe, Eiji Takimoto, Kazuyuki Amano and Akira Maruoka [1]
Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan

1 Learning via Random Projection

Random projection is a technique of mapping a number of points in a high-dimensional space into a low dimensional space with the property that the Euclidean distance of any two points is approximately preserved through the projection. A random projection from $\mathbf{R}^n$ to $\mathbf{R}^k$ is typically defined as a random $n \times k$ matrix $R = (r_{ij})$ with each entry chosen randomly and independently according to some probability distribution: an n-dimensional vector $u \in \mathbf{R}^n$ is projected to $u' = \frac{1}{\sqrt{k}} R^T u = \frac{1}{\sqrt{k}} (\sum_{i=1}^n r_{i1} u_i, \ldots, \sum_{i=1}^n r_{ik} u_i)$. Here a constant $1/\sqrt{k}$ is a normalization factor which makes the expectations of the Euclidean norm of u' and u coincide, i.e., $\mathbf{E}[||u||] = \mathbf{E}[||u'||]$.

The development of the random projection was originated by Johnson and Lindenstrauss [5] who showed that such embeddings are in fact possible. Recently, several simple projections were proposed and successfully applied for designing good learning algorithms for various concept classes (e.g., [1, 3, 6]).

Let $N(0,1)$ denote the standard normal distribution with mean 0 and variance 1, and $U(-1,1)$ denote the discrete distribution defined by $r = 1$ with probability $1/2$ and $r = -1$ with probability $1/2$.

Theorem 1 *[3] Let $R = (r_{ij})$ be a random $n \times k$ matrix, such that each entry r_{ij} is chosen independently from either $N(0,1)$ or $U(-1,1)$. Then for any fixed vector $u \in \mathbf{R}^n$ and any $\epsilon > 0$, the vector $u' = \frac{1}{\sqrt{k}} R^T u$ satisfies*

$$\Pr[\big|||u'||^2 - ||u||^2\big| \geq \epsilon ||u||^2] \leq 2e^{-\epsilon^2 k/8}.$$

Suppose that we are given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ consisting of m pairs of an example in $\mathbf{R}^n$ and its label $y_i \in \{-1, 1\}$, where each pair (x_i, y_i) in S is chosen independently according to a distribution D on $\mathbf{R}^n \times \{-1, 1\}$. Suppose further that S can be correctly classified by a hyperplane with margin l, i.e., there exists a $h \in \mathbf{R}^n$ with $||h|| = 1$ such that for every $1 \leq i \leq m$, $y_i(h^T x_i) \geq l$ holds.

Corollary 1 *Let $L = \max_i ||x_i||$ and $k = O(\frac{72L}{l^2} \ln \frac{6m}{\delta})$. Let $S' = \{(x'_1, y_1), \ldots, (x'_m, y_m)\}$ be the projected sample by an $n \times k$ matrix R defined as above, i.e., $x'_i = (1/\sqrt{k}) R^T x_i$ and h be the projected classifier, i.e., $h' = (1/\sqrt{k}) R^T h$. Then the probability that h' doesn't correctly classify S is at most δ.*

Corollary 1 guarantees that the following algorithm can learn an n-dimensional hyperplane of margin l in a PAC sense if we set k to $O(\frac{1}{l^2} \ln \frac{1}{\epsilon l \delta})$ and m to $O(\frac{1}{l^2 \epsilon} \ln \frac{1}{\epsilon} \ln \frac{1}{\epsilon l \delta})$ [3].

Hyper-Plane Algorithm

1. Choose an $n \times k$ random matrix R by picking each entry independently from $U(-1,1)$ or $N(0,1)$.
2. Obtain m examples from D and project them to $\mathbf{R}^k$ using R.
3. Run the Perceptron Algorithm in $\mathbf{R}^k$: Let $w = 0$. Perform the following until all examples are correctly classified: Pick an arbitrary misclassified example $(x'_i, y_i) \in \mathbf{R}^k \times \{-1, 1\}$, let $w = w + y_i x'_i$.
4. Output $f(x) = \mathrm{sign}(w^T (R^T x))$ as a final hypothesis.

[1] E-mail: {nabe,t2,ama,maruoka}@maruoka.ecei.tohoku.ac.jp

2 Random Projection with Limited Independence

By inspecting the proof of Theorem 1 carefully, we can show that the i.i.d condition of the entries of a projection matrix R can be weakened at the cost of slighly loosing the confidence of length preserving.

Definition 1 *Let k, n be two integers such that $k \leq n$. Let D be a probability distribution on $\{-1,1\}^n$. An n-vector of random variables $x = (x_1, \ldots, x_n) \in \{-1,1\}^n$ with respect to D is said to be k-independent, if $\forall j \leq k$, $\forall I = \{i_1, \ldots, i_j\} \subseteq \{1, \ldots, n\}$ and $\forall V = (v_1, \ldots, v_j) \in \{-1,1\}^j$,*

$$\Pr_D[x_I = V] \quad = \quad \prod_{l=1}^{j} \Pr[x_{i_l} = v_l] = \frac{1}{2^j}$$

holds. Here $x_I = V$ denotes $(x_{i_1} = v_1) \wedge \cdots \wedge (x_{i_j} = v_j)$.

It is well known that n random variables with k-wise independence can be efficiently generated using only $O(k \log n)$ random bits (e.g., [2]). By observing that the expectation and the variance of the norm of the projected vector can be expressed by the 2^{nd} and the 4^{th} moments of the random variables r_{ij}, we can show the following result by using Chebyshev's inequality (see [7] for the proof).

Theorem 2 *Let $u \in \mathbf{R}^n$ be arbitrary but fixed. Let $R = (r_{ij})$ be a random $n \times k$ matrix such that each entry r_{ij} is chosen according to $U(-1,1)$ and for every $1 \leq i \leq k$, the i-th column of R, i.e., $R_i = (r_{1i}, \ldots, r_{ni})$ is 4-wise independent. Then for every $\epsilon > 0$, the vector $u' = \frac{1}{\sqrt{k}}(R^T u)$ satisfies*

$$\Pr[\left| ||u'||^2 - ||u||^2 \right| \geq \epsilon ||u||^2] \leq \frac{2}{\epsilon^2 k}.$$

The above theorem guarantees that a 4-wise independent distribution has a desired property although its performance guarantee is weaker than that of the i.i.d case (Theorem 1). It is worthwhile to note that the 4-wiseness in the theorem is optimal in the following sense:

Theorem 3 *There exists a 3-wise independent distribution D that satisfies the following: if each column of R is chosen independently according to D, then for every k, the vector $u' \in \mathbf{R}^k$ projected by R from $u = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1) \in \mathbf{R}^n$ (i.e., $||u|| = 1$) satisfies*

$$\Pr\left[||u'|| = \frac{1}{n} \right] = \left(1 - \frac{1}{n}\right)^k > \frac{1}{2e}.$$

In order to construct an efficient learning algorithm for a hyperplane using a random projection with bounded independence (i.e., via Theorem 2 instead of Theorem 1), it seems to be necessary to improve the upper bounds on the error probability in Theorem 2 to an exponentially small in k as in Theorem 1. We believe that such improvements would be possible by a more refined analysis with combination of Chebyshev's inequality and Azuma's inequality.

References

[1] D. Achlioptas: Database Friendly Random Projections, Proc. 20th PODS, pp. 274–281, 2001.

[2] N. Alon, L. Babai and A. Itai, A Fast and Simple Randomized Parallel Algorithm for the Maximal Independent Set Problem, J. Algorithms, 7, pp. 567–583, 1986.

[3] R.I. Arriaga and S. Vempala: An Algorithmic Theory of Learning: Robust Concepts and Random Projection, Proc. 40th FOCS, pp. 616–623, 1999.

[4] P. Indyk and R. Motowani: Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, Proc. 30th STOC, pp. 604–613, 1998.

[5] W.B. Johnson and J. Lindenstrauss: Extensions of Lipschits Mapping into a Hilbert Space, Contemp. Math., 26, pp. 189–206, 1984.

[6] A.R. Klivans and R.A. Servedio: Learning Intersections of Halfspaces with a Margin, Proc. 17th COLT, pp. 348–362, 2004.

[7] T. Watanabe, E. Takimoto and A. Maruoka, Dimensionality Reduction by Random Projection (in Japanese), Tech. Rep. of IEICE, COMP 2001-92, 2002.