Clustering of Data and Nearest Neighbors Search for Pattern Recognition with Dimensionality Reduction Using Random Projections

Ewa Skubalska-Rafajłowicz

Institute of Computer Engineering, Control and Robotics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50–370 Wrocław, Poland
ewa.rafajlowicz@pwr.wroc.pl

Abstract. The dimensionality and the amount of data that need to be processed when intensive data streams are classified may occur prohibitively large. The aim of this paper is to analyze Johnson-Lindenstrauss type random projections as an approach to dimensionality reduction in pattern classification based on K-nearest neighbors search. We show that in-class data clustering allows us to retain accuracy recognition rates obtained in the original high-dimensional space also after transformation to a lower dimension.

1 Introduction

In this paper we develop and examine some new properties of the random projections when applied to pattern recognition algorithms based on nearest neighbors search.

The nearest neighbors search problem is of major and growing importance to a variety of applications such as data compression, statistical data analysis, pattern recognition, data mining, etc [3]. In many applications (image or video processing, analysis of data from sensors arrays, CCD and CMOS cameras, speech recordings and spectral analysis) the dimensionality of the feature space can attain hundreds or even several thousands.

Several dimension reduction techniques, linear and non-linear, have been proposed in many fields, including principal components analysis, projection pursuit, principal curves, independent component analysis, self-organizing maps, space-filling curves [14], neural network implementations of some statistical models and random linear projections, among others. It is reported in the literature that random projections work better than other simplistic dimension reduction schemes [4] when original dimensionality of the data is very large.

In the random projection method, the original high-dimensional observations are projected onto a lower-dimensional space using a suitably scaled random matrix with independent, typically, normally distributed entries.

L. Rutkowski et al. (Eds.): ICAISC 2010, Part I, LNAI 6113, pp. 462–470, 2010.
© Springer-Verlag Berlin Heidelberg 2010

Random projections have been found to be a computationally efficient, yet sufficiently accurate method for dimensionality reduction. Promising experimental results are reported in many papers [3], [5], [7], [13], [15]. There has been an increasing interest in avoiding the curse of dimensionality by resorting to approximate nearest neighbor searching [8], [9] and clustering methods [3], [4]. However, in pattern recognition methods based on voting among K-nearest neighbors, the true class-membership of the K-nearest neighbors is of major importance. Thus, we propose the method of verifying when the nearest neighbor found in lower-dimensional space (space obtained by random projections) is with high probability the true nearest neighbor (the nearest neighbor in the original high-dimensional space). On this bases we propose the adaptive method of neighbors number selection by voting among neighbors. Furthermore, we examine the role of data compression, showing that in-class data clustering allows us to retain accuracy recognition rates (obtained in the original high-dimensional space) after projection to lower-dimensional space.

2 Johnson-Lindenstrauss Lemma and Linear Random Projections

In random projections, we can estimate the original pairwise Euclidean distances directly using the corresponding Euclidean distances in smaller dimension. Furthermore, the Johnson-Lindenstrauss lemma [11], [6] provides the performance guarantee.

The Johnson-Lindenstrauss Lemma can be formulated as follows: A d-dimensional point set C can be embedded into a k-dimensional space, where k is logarithmic in cardinality of C and independent of d. Distances between points in C are preserved within a factor $(1 \pm \varepsilon)$, i.e., there exists a mapping $F : R^d \rightarrow R^k$ such that for all $x, y \in C$

$$(1-\varepsilon)||x-y||^2 \leq ||F(x)-F(y)||^2 \leq (1+\varepsilon)||x-y||^2. \tag{1}$$

A random projection from d dimensions to k dimensions is a linear transformation represented by a $d \times k$ matrix $S \in R^{k\times\ d}$ – a matrix whose entries are i.i.d. samples of some random variable. Let $u \in R^d$ be an original data point. The projected data is a point in k -dimensional space (R^k) given by $v = S.u$

We can estimate the distance in d-dimensional space $||u_m - u_l||^2$ from the sample squared distances as follows:

$$\hat{D}^2 = \frac{1}{k}\sum_{j=1}^{k}(v_{mj} - v_{lj})^2 = ||v_m - v_l||^2. \tag{2}$$

Note that $\hat{D}^2$ is estimated using the elements of projected vectors and every projection contains some information about the length of the projected vector (u_m, u_l).

It is easy to see that $E\{\hat{D}^2\} = D^2$, while its variance, decreases to zero as $k \to \infty$, since $var(\hat{D}^2) = \frac{2}{k}D^4$.

Using chi-squared tail Chernoff bounds (see [6] for details) we can obtain the bound on the probability when the relative error exceeds ε $(1 > \varepsilon > 0)$

$$Pr\left\{\frac{|\hat{D}^2 - D^2|}{D^2} \geq \varepsilon\right\} \leq 2\ \exp\left(-\frac{k}{4}\varepsilon^2 + \frac{k}{6}\varepsilon^3\right). \tag{3}$$

The Bonferroni union bound leads to a very rough (too conservative) approximation of k, because it ignores the correlations. In practice, one can use much smaller values of k with acceptable distance distortion (see [3],[5], [7], [12], [13], [15] among others).

Table 1. Comparison of M-NN method and probable M-NN method for two Gaussians problems

		Example A		Example B		Example C	
M	k	M-NN	P M-NN	M-NN	P M-NN	M-NN	P M-NN
1	1000	1.00	1.0-0.99	1.00	0.94 - 0.74	0.73	0.69-0.49
1	300	1.0	1.0-0.96	1.00	0.83-0.69	0.88	0.63-0.48
1	100	1.0	0.95-0.80	1.00	0.76-0.56	0.74	0.63-0.45
10	1000	1.0	1.0-0.99	1.00	1.0-0.94	0.88	0.72-0.54
10	300	1.0	1.0-0.99	1.00	0.95-0.78	0.87	0.65-0.48
10	100	1.0	1.0-0.94	1.00	0.84-0.57	0.87	0.71-0.46

Table 2. Accuracy rates using PM-NN method for AM problem (left table) and BM (right table) for two mixtures of Gaussian problems

			Example AM	
N	M	k	mean	range
10	1	100	0.87	0.9-0.79
10	3	100	0.93	0.97-0.81
10	5	100	0.93	0.97-0.87
100	1	100	0.944	0.99-0.91
100	3	100	0.971	1.0-0.93
100	5	100	0.98	1.0-0.94

			Example BM		
N	M	k	M-NN	mean	range
100	1	100	0.932	0.572	0.69-0.48
100	3	100	0.975	0.577	0.64-0.45
100	5	100	0.99	0.579	0.66-0.47
100	1	300	0.930	0.61	0.70-0.52
100	3	300	0.973	0.66	0.72-0.55
100	5	300	0.986	0.67	0.72-0.61
100	1	1000	0.944	0.688	0.77-0.60
100	3	1000	0.979	0.735	0.83-0.65
100	5	1000	0.988	0.765	0.85-0.71

Table 3. Accuracy rates using probable M-NN method for AM problem (left table) and BM problem (right table)with adaptively chosen M (two mixtures of Gaussian problems)

ε	$M \geq$	k	Example AM mean	range
0.02	1	100	0.964	0.99-0.89
0.01	1	100	0.960	0.99-0.91
0.005	1	100	0.945	0.95-0.89
0.01	3	100	0.975	1.00-0.93
0.01	5	100	0.97	1.00-0.92

ε	$M \geq$	k	Example BM M-NN mean		range
0.01	1	100	0.957	0.585	0.697-0.47
0.1	1	100	0.957	0.578	0.67-0.48
0.01	3	100	0.99	0.595	0.67-0.48
0.01	5	100	0.993	0.615	0.73-0.49
0.01	1	300	0.973	0.655	0.73-0.60
0.01	3	300	0.986	0.666	0.75-0.61
0.01	5	300	0.944	0.684	0.77-0.57
0.01	1	1000	0.979	0.715	0.72-0.67
0.01	3	1000	0.988	0.762	0.85-0.68
0.01	5	1000	0.998	0.789	0.86-0.61

Table 4. Accuracy rates obtained using probable M-NN method for AM problem (left table) and BM problem (right table) with centers of Gaussian as class prototypes

ε	$M \geq$	k	Example AM mean	range
0.01	1	100	0.996	1.00-0.98
0.01	3	100	0.997	1.00-0.98
0.01	5	100	0.996	1.00-0.98
0.01	1	300	1.00	1.00-1.00
0.01	3	300	1.00	1.00-1.00
0.01	5	300	1.00	1.00-1.00

ε	$M \geq$	k	Example BM mean	range
0.01	1	100	0.776	0.84-0.68
0.01	3	100	0.802	0.85-0.74
0.01	5	100	0.804	0.86-0.74
0.01	1	300	0.877	0.92-0.84
0.01	3	300	0.915	0.96-0.84
0.01	5	300	0.912	0.95-0.85
0.01	1	1000	0.985	1.00-0.97
0.01	3	1000	0.992	1.00-0.97
0.01	5	1000	0.970	1.00-0.94

Table 5. Accuracy rates obtained using probable M-NN method for AM problem (left table) and BM problem (right table) with class means as prototypes

		Example AM	
$M \geq$	k	mean	range
1	10	0.809	0.92-0.76
1	50	0.971	1.00-0.93
1	100	0.995	1.00-0.98
1	300	1.00	1.00-0.99
1	1000	1.00	1.00-0.99

		Example BM		
M	k	1-NN	mean	range
1	100	1.00	0.796	0.84-0.73
1	300	1.00	0.89	0.95-0.83
1	1000	1.00	0.977	0.99-0.95

3 Probable Nearest Neighbors - Preservation of Closeness Relation under Space Embedding Using Normal Random Projections

Let $x \in X$ be a given query point. The nearest neighbors search problem is defined as follows. Given a set of M points $C = \{c_1, c_2, \ldots, c_M\}$ in a metric space (X, d). Order the points in C in such a way that $c_{(j)}$ is the j -th closest point among C to a query point $x \in X$.Without loss of generality we can assume that $D_i(x) \leq D_{i+1}(x)$, $i = 1, \ldots, M-1$, where $D_i(x) = d(c_i, x)$. Thus $D_i(x)$ is the i-th nearest neighbor of x among the points from set C, $|C| = M$.

According to Johnson-Lindenstrauss lemma

$$Pr\{1 - \varepsilon \leq \frac{\hat{D}_i^2(x)}{D_i^2(x)} \leq 1 + \varepsilon, i = 1, \ldots, M\} \geq 1 - 2M \exp(-\frac{k}{4}\varepsilon^2 + \frac{k}{6}\varepsilon^3). \quad (4)$$

It is easy to see that with probability exceeding the right side of 4 we have the following result. If $\varepsilon \leq \frac{D_j^2(x) - D_i^2(x)}{D_j^2(x) + D_i^2(x)}$ then

$$(1 - \varepsilon) D_i^2(x) \leq \hat{D}_i^2(x) \leq \hat{D}_j^2(x) \leq (1 + \varepsilon) D_j^2(x), \quad (5)$$

for any $i, j = 1, \ldots, M$ and $i < j$. This means, that if squared distances from x to neighbors c_i and c_j differ at least $\varepsilon(D_j^2(x) + D_i^2(x))$, then $\hat{D}_i^2(x) \leq \hat{D}_j^2(x)$. It is obvious, that we want to know when we can conclude from $\hat{D}_i^2(x) \leq \hat{D}_j^2(x)$ that the original squared distances are in the same relation: $D_i^2(x) \leq D_j^2(x)$ with probability at least $1 - \delta$, $\delta \in (0, 1)$.

Note that for every $i = 1, \ldots, M$

$$Pr\{\frac{\hat{D}_i^2(x)}{D_i^2(x)} \geq 1 + \varepsilon\} = Pr\{\frac{D_i^2(x)}{\hat{D}_i^2(x)} \leq \frac{1}{1 + \varepsilon}\} \geq Pr\{\frac{D_i^2(x)}{\hat{D}_i^2(x)} \leq 1 - \varepsilon\}.$$

Similarly,

$$Pr\{\frac{\hat{D}_i^2(x)}{D_i^2(x)} \leq 1-\varepsilon\} = Pr\{\frac{D_i^2(x)}{\hat{D}_i^2(x)} \geq \frac{1}{1-\varepsilon}\} \geq Pr\{\frac{D_i^2(x)}{\hat{D}_i^2(x)} \geq 1+2\varepsilon\},$$

provided that $\varepsilon \leq 1/2$.

Thus,

$$Pr\{1-\varepsilon \leq \frac{D_i^2(x)}{\hat{D}_i^2(x)} \leq 1+2\varepsilon, i=1,\ldots,M\} \geq$$

$$Pr\{1-\varepsilon \leq \frac{\hat{D}_i^2(x)}{D_i^2(x)} \leq 1+\varepsilon, i=1,\ldots,M\} \geq 1-2M\exp(-\frac{k}{4}\varepsilon^2+\frac{k}{6}\varepsilon^3),\ \varepsilon \in (0,1/2].$$

So, if $\hat{D}_i^2(x) < \hat{D}_j^2(x)$, and the relative distance between two neighbors is

$$\frac{\hat{D}_j^2(x) - \hat{D}_i^2(x)}{\hat{D}_j^2(x) + 2\hat{D}_i^2(x)} \geq \varepsilon, \tag{6}$$

then with probability exceeding $1-2M\exp(-\frac{k}{4}\varepsilon^2+\frac{k}{6}\varepsilon^3)$ we obtain $D_j^2(x) \geq (1-\varepsilon)\hat{D}_j^2(x) \geq (1+2\varepsilon)\hat{D}_i^2(x) \geq D_i^2(x)$, for any $i,j \in \{1,\ldots,M\}$.

Now we can easily prove the following theorem.

Theorem 1. *Let $\hat{D}_1(x) < \ldots < \hat{D}_M(x)$. Choose $\delta > 0$ and $\varepsilon \in (0,1/2)$ such that $\varepsilon < \frac{\hat{D}_j^2(x)-\hat{D}_i^2(x)}{\hat{D}_j^2(x)+2\hat{D}_i^2(x)}, i,j=1,\ldots,M,\ i<j$. If the dimensionality after reduction k is selected in such a way that*

$$k \geq 4\frac{\log 2 + \log M - \log\delta}{\varepsilon^2 - \frac{2}{3}\varepsilon^3}, \tag{7}$$

then the probability that for any pair (i,j), $D_i(x) > D_j(x)$ $(i,j=1,\ldots,M, i<j)$ is smaller than δ.

Note that this theorem is formulated for only one query point. However, as in the case of the probabilistic version of Johnson-Lindenstrauss theorem, Theorem 1 is also too restrictive for practical purposes. Nevertheless, we can use inequality (6) for adaptive selection of a number of nearest neighbors. We propose the following algorithm:

Adaptive selection of number of neighbors: Let x be a query point. Assume that we have chosen M nearest neighbors of x from a given prototype set P (consisting of $N >> M$ elements). Choose $\varepsilon > 0$.

Step 1. Find $(M+1)$-th nearest neighbor of x from set P and check inequality (6) for $i = M$ and $j = M+1$.

Step 2. If (6) fails, set $M = M+1$ and go to Step 1. Otherwise, end the procedure.

4 Experimental Results

In the first part of the experiments, we have used the following three data sets. Two equiprobable classes described by normal distributions with vector means: $(0, 0, \ldots, 0)$ and $(m, m, \ldots, m)$, respectively, and the same covariance matrix I. Dimension of the original data d is equal to $d = 10000$ and $m = 0.5$ (set A with Bayes error close to 0), $m = 0.25$ (set B with Bayes error also close to 0) and $m = 0.05$ (set C with Bayes error 10^{-12}).

We have compared the performance of the M-NN nearest neighbor method based on data taken from the original space and the probable M-NN (PM-NN) method based on projected data(using normal LRP). M denotes the number of neighbors ($M = 1$ or $M = 10$). The length of a learning sequence N was fix as 100. Note that it is very small learning sequence in comparison to the dimensionality of the space. The dimension after projection k was established as 1000, 300 and 100. In each case we run 20 experiments with different random projections. Table 1 contains the accuracy rates (accuracy is 1 minus error rate) for both methods.

Note that in each classification problem the centers of two Gaussians are $m\sqrt{d}$ apart. It is known, that two Gaussian $N(\mu_1, I_d)$ and $N(\mu_2, I_d)$ are c-separated if $||mu_1 - \mu_2|| \geq c\sqrt{d}$. In high dimensional spaces if c is between $0.5 - 1$ then Gaussians have an almost negligible overlap. Thus, only for $c = 0.05$ (Example C) larger recognition errors occur.

In the second part of the experiments, we concentrate on more complicated problems. Namely, a class-distribution for each of two equiprobable classes consists of a mixture of five Gaussians (with the same covariance matrix I) and concentrated around $(0, 0, \ldots, 0)$ and $(m, m, \ldots, m)$, respectively.The dimension of the original data d is equal to $d = 10000$ and $m = 0.5$ (problem AM – almost negligible overlap), $m = 0.1$ (problem BM – slightly overlapping classes). In both cases Bayes error is close to 0.000. N denotes the length of the learning sequence. The results are summarized in Table 2.

The same problems have been further examined using adaptively chosen number of nearest neighbors.The results are given in Table 3. Table 4 contains accuracy rates for decisions made on the basis of distances of query points to $N = 10$ class prototypes (five for each class). We assume that prototypes are obtained using some kind of cluster formation [7], [13], [12]. For the sake of simplicity and in order to avoid cross-interferences between methods, prototypes are taken as centers of Gaussians forming the mixtures. Similar experiments have been performed for 2 class-prototypes (one for each class). This time, the approximate distribution means $(0, 0, \ldots, 0)$ and $(m, m, \ldots, m)$ form the prototypes set.

5 Conclusions

Random projections methods outperform other dimensionality reduction methods from the view point of computational complexity. If data clusters are only slightly mixed, then $1 - NN$ recognition error grows very rapidly, when a dimension of the projection space decreases, since clusters after projection are

strongly mixed. Thus, using the many neighbors approach with an adaptively chosen number of neighbors allows us to improve the classification efficiency. Furthermore, in many cases, carefully designed class prototypes enable us to retain almost the same level of recognition accuracy even if dimensionality reduction is very large.

In this paper we have analyzed some aspects of Euclidean nearest neighbors classification methods. Inner product based distance measures should be investigated in a similar vein. The results presented in [16], [12] can be a good starting point.

Acknowledgments. This work was supported by a grant of the Ministry of Science and Higher Education for 2006-2009.

References

1. Achlioptas, D.: Database friendly random projections. In: Proc. Principles of Database Systems (PODS), pp. 274–281 (2001)
2. Arriaga, R., Vempala, S.: An algorithmic theory of learning: Robust concepts and random projection. In: Proc. of FOCS, New York, pp. 616–623 (1999)
3. Bertoni, A., Valentini, G.: Random projections for assessing gene expression cluster stability. In: Proceedings IEEE international joint conference on neural networks, vol. 1, pp. 149–154 (2005)
4. Biau, G., Devroye, L., Lugosi, G.: On the Performance of Clustering in Hilbert Spaces. IEEE Tran. on Information Theory 54(2), 781–790 (2008)
5. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proc. Seventh ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD 2001), pp. 245–250 (2001)
6. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures and Algorithms 22(1), 60–65 (2003)
7. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of the 20th international conference on machine learning (ICML 2003), Washington DC, USA (August 2003)
8. Indyk, P., Motwani, R.: Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proc. of STOC, Dallas, TX, pp. 604–613 (1998)
9. Indyk, P., Naor, A.: Nearest neighbor preserving embeddings. ACM Transactions on Algorithms (TALG) archive 3, article no. 31 (2007)
10. Clarkson, K.: Nearest-neighbor searching and metric space dimensions. In: Nearest-Neighbor Methods for Learning and Vision: Theory and Practice. MIT Press, Cambridge (2005)
11. Johnson, W.B., Lindenstrauss, J.: Extensions of Lipshitz mapping into Hilbert space. Contemporary Mathematics 26, 189–206 (1984)
12. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering 18, 92–106 (2006)
13. Kaski, S.: Dimensionality reduction by random mapping: fast similarity computation for clustering. In: Proc. IEEE International Joint Conference on Neural Networks, vol. 1, pp. 413–418 (1998)

14. Skubalska-Rafajlowicz, E.: Pattern recognition algorithm based on space-filling curves and orthogonal expansion. IEEE Trans. on Information Theory 47, 1915–1927 (2001)
15. Skubalska-Rafajlowicz, E.: Random projection RBF nets for multidimensional density estimation. International Journal of Applied Mathematics and Computer Science 18(4), 455–464 (2008)
16. Skubalska-Rafajlowicz E: Neural networks with sigmoidal activation functions – dimension reduction using normal random projection. Nonlinear Analysis 71(12), e1255–e1263 (2009)
17. Vempala, S.: The Random Projection Method. American Mathematical Society, Providence (2004)