Empirical Determination of Lower Bounds on RP Embedding

Lili He and Ian R. Greenshields

Department of Computer Science and Engineering
University of Connecticut
Email: lili@engr.uconn.edu and ian@engr.uconn.edu

Abstract

Data analysis is critical to many (if not all) Homeland Security missions. Data to be fielded in this domain is typically immense in cardinality (number of records) and immense in dimensionality (features per record). Random Projections (RP) have been proposed as an effective technique for embedding a metric space of dimension d into one of dimension k while retaining bounds on the distortion of the metric under embedding. More recently, Achlioptas has shown from work founded on the Johnson-Lindenstrauss lemma on embeddings that a sparse random matrix could give additional computational savings in random projection, and he also gives a theorical embedding bound k_0. However, empirical evidence shows that one can choose $k \ll k_0$ and yet retain low distortion in the metric. In the paper we propose an experimental approach to find the optimum lower bound of k_0 for any distribution models in terms of classification.

1. Introduction

As one of the key features of many homeland security initiatives, data analysis is widely involved in discovering unknown, valid patterns and relationships in large data sets to, for example, identify and track terrorist activities and individual terrorists themselves and also in statistics concerning components of the system. Data clustering is itself a massive subject with an enormous literature, and clustering and classification in this context may be required to be secondary to some form of data reduction in order to bring any one of the various techniques open to an analyst into computational realizability. Concomitant to this is the need to retain clustering fidelity in the reduced dimensional space relative to clusters in the original data expressed in the original high dimensional space. Numerous techniques are open to the analyst to accomplish this dimensionality reduction, including such obvious techniques as SVD, FFT, DCT or DWT. Each of these techniques has benefits (and drawbacks) relative to any particular problem at hand. More recently, Random Projections have been proposed as an alternative dimensionality reducing technique, and there is evidence that they are both computationally efficient and valuable. The formal intent of a random projection is to embed a metric space of dimension d (the features of the dataset) into one of dimension $k < d$ while retaining bounds on the distortion of the metric under the embedding. Achlioptas [1] has recently shown (from work founded on the Johnson-Lindenstrauss theorem [2] on embeddings) that a set Q of n points in R^d can be embedded into R^k for $k \geq k_0$ (using a sparse random matrix leading to an efficient implementation of the projection) where

$$k_0 = (4+2b)\cdot \log n/(e^2/2 - e^3/3) \qquad (1)$$

Here the parameters e and b control distortion in the metric and probability that this distortion will be met, and we have additionally that the distortion is bounded above and below (up to b) by

$$(1-e)\,||u-v||^2 \leq ||P_u - P_v||^2 \leq (1+e)\,||u-v||^2 \qquad (2)$$

for $u, v \in Q$ under projection P. Note that the dimension k of the embedding (up to the bounds of the theorem) is proportional to the logarithm of the cardinality of the number of points to be embedded into R^k from R^d. In very large datasets the theoretical value of k (for given e and b) may be quite large. However, *empirical* evidence shows that one can choose $k \ll k_0$ and yet retain low distortion in the metric. More formal evidence that data from a mixture of m Gaussians can be projected into just $O(\log m)$ dimensions while still retaining the approximate level of separation between the clusters is given in [3]. This claim is significant for certain applications. Unfortunately it is not suitable for non-Gaussian mixture data, such as text data, whose term frequency matrix is sparse (e.g., having Poisson character). In the paper we propose an experimental approach to find the optimum lower bound of k_0 for any distribution models in terms of classification.

2. Methodology

Our goal is to explore a variety of Random Projections into k to uncover both a suitable examplar of the projection and k itself. Here k is deemed acceptable (or not) based on a simple metric distortion

$$e = abs(||P_u - P_v|| - ||u-v||) \,/\, ||u-v|| \qquad (3)$$

Evidently, one wishes to avoid the explicit projection of the entire population dataset to determine these factors. Efficient and accurate data sampling is of importance. The most commonly used data sampling techniques in statistics is random sampling, where the selection of a sample from a population is based on the principle of randomization and every element has a chance of being selected. However, it isn't very suitable for our application, since the optimum

size of sample data is unknown and tends to be subjective in some applications. In addition, it will introduce additional randomness. We propose, in the context of classification, to use *marginal points* as proxies for samples within the dataset, and define the *marginal points* as the *support vectors (SVs)* in the original dimensional space.

Support vectors [4], as shown in figure 1, in essence, are the ones closest to the decision boundary and are of significance in classification, which means if we remove all other training points, and repeat the training, the same hyperplane will be found. For a given data set, the number of SVs may increase when the dimension decreases, while SVs in high dimensional space will still be the *marginal points* in the low dimensional space. Moreover, since we expect the number of the support vectors in original space to be small compared to a larger subset of training points we retain efficiency where needed. The proposed procedure of finding optimum lower bound of k_0 is shown in Table 1.

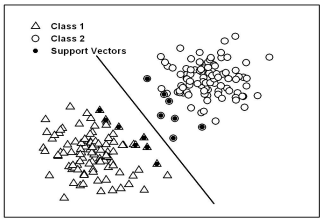

Figure 1. Support vectors in linear separable case.

3. Experiments and Results

We designed experiments to demonstrate: 1) the procedure of finding lower bound of k_0 outlined in Table 1 works well; 2) Sampling data set with marginal points outperforms random sampling. We used four linearly separable data sets, as summarized in table 2. Two text data sets are from (http://www.cs.cmu.edu/~textlearning). The other two are synthetic data sets. Text data are represented in a vector space, in which each document forms a d -dimensional vector, where d is the vocabulary size. Elements of the vector indicate term frequencies. The stop words were removed from the documents.

3. 1 Finding optimum lower bound of k_0

The behavior of the process of finding lower bound of k_0 on SYN1 is exemplified in Figure 2. The threshold d was set to 0.06, which is subjective to be fine-tuned. To further validate the results, SVM were implemented to classify the data sets. The parameters of classifier were trained using randomly picked (70%) data points from each class, while the rest of (30%) of the data were used to test. The results are enumerated in Table 3.

SVM achieved over 90% accuracy rates for Sci and Politics

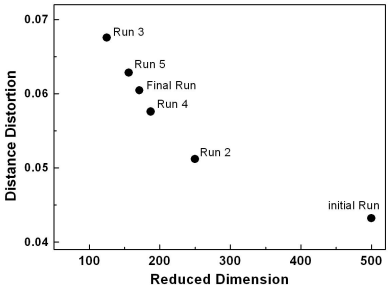

Figure 2. The process of finding lower bound of k_0 on SYN1.

Table 1 Finding optimum lower bound of k_0 .

Input: $A_{n\times d}$,where n is the number of points, d is the original dimension.
Output: the optimum lower bound of k_0 .
1. Run SVM to select support vectors (marginal points in following reduced dimensions).
2. Initialize $k = \lfloor d/2 \rfloor$, d ,*head* = *0*, and *tail* = *d*.
3. $E_{n\times k} = random\text{Pr}ojection(A,k)$.
4. Compute averaged pair-wise distance distortion e among the marginal points.
Repeat
  If $e < d$ then
    tail =*k*; *k* = floor (*head* + (*tail* – *head*) / *2*);
    Otherwise
    head= *k*; k = floor(*head* + (*tail* – *head*) / *2*);
  End if
  $E = random\text{Pr}ojection(A,k)$
  Compute e among the marginal points.
Until $e \approx d$
5. Return k as k_0

Table 2. Summary of the testing data sets.

Data Sets	SCI (med & space)	Politic (gun & mideast)	SYN1 (synthetic Data)	SYN2 (synthetic Data)
# of points	2000	2000	4000	3000
# of class	2	2	4	3
d	5000	1000	1000	1000

Table 3. Summary of empirical found lower bounds of k_0

Data	Sci	Politics	SYN1	SYN2
Empirical k_0	85	86	179	171
Classificat ion rate at Empirical k_0	91.5%	94.2%	90.3%	93.5%
e calculated among marginal points	0.061	0.058	0.060	0.061
e calculated among random points	0.065	0.059	0.599	0.485

data sets at dimension about 90 and for synthetic data at dimension about 180, respectively. Comparing to the theoretical k_0 , which would be on the order of 1000 for the data set of 2000 points and $e < 0.4$. Thus, we may conclude that our approach is able to find significantly lower dimension than both the original space and expected embedding space, where the metric is still preserved well so

that the data may therefore be possible to carry almost the same metric information into a much lower dimensional space than expected via random projection.

3.2 Marginal points sampling outperforms random sampling

First of all, for a given data set, the number of SVs in original dimension is fixed so that the sampling size is evidently known. Secondly, the variance of marginal points sampling is lower than random sampling, where distance distortions will vary a lot for different sampling even without consider of the effect of random projection. As shown in Figs 3 and 4, the averaged variance of the distance distortion in marginal points sampling is ± 0.0083, while it is ± 0.0384 in random sampling.

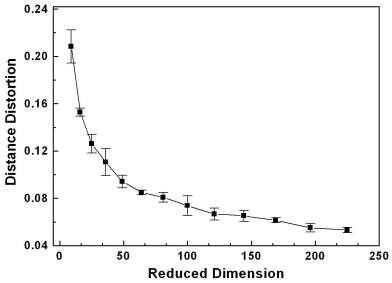

Figure 3. Variance of *e* in marginal points sampling on SYN1.

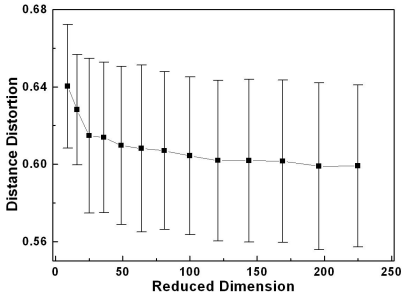

Figure 4. Variance of *e* in random sampling on SYN1.

Finally, marginal points sampling will favor us to preset threshold of distance distortion d. We expect the same threshold with a small variance could be used for different data sets. For the purpose of comparing, the same number of marginal points and random instances are used to compute distance distortion *e*, as shown in Table 3. Not surprisingly, we find that the distortions calculated among marginal points are concentrated to a particular value (0.06) for all four data sets, which may be a good candidate for *universal* and *concentrated* threshold. But in random sampling, those of two synthetic data sets are almost 10 times larger than those of the text data sets, while the classification rates are still as high as 90.3% and 93.5%. The reason of this may be that the marginal points are of importance to the classification, the separability of a data set is nearly only affected by their distance distortions at each reduced dimension. However, this is not the case for arbitrary randomly picked instances, which may be from the body of a distribution for each class or may be only from one of several classes so that the value of distance distortions of these instances may not directly relate to classification results. Thus, we conclude that marginal points sampling makes more sense than random sampling in the context of classification.

4. Conclusions

In the paper, we introduced the problem of the empirical lower bound of k_0 would be much lower than the theoretical one in random projection, and brought forward an algorithm to find that optimum lower bound of for any distribution models in terms of classification. We proposed to use marginal points sampling in the context of classification. Experimental results showed that our approach successfully found a significantly lower dimension for projections than the theoretical one, where the metric is still preserved well so that it is possible to carry almost the same metric information as that in the original space. Moreover, we designed experiments to show that marginal points sampling is better than random sampling, since it could specifically and efficiently determine the sample size. The lower variance of marginal points sampling than random sampling makes the result of our algorithm more accurate.

The knowledge of how low of a dimension that a data set could be reduced to is of very importance to efficient and accurate data analysis. This paper only takes a small step toward that direction—the precondition is that the labels of the data points are known. Other worthwhile future work includes how to find the optimum lower bound for data sets without labels as well as a thorough theoretical analysis of how tight the bound of the random projection could be.

This work was supported by a grant from DARPA.

References

[1] Achilioptas, D. 2003. Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins. *Journal of Computer and System Sciences* 66(4):671-687.

[2] Frankl, P.; and Maehara, H. 1988. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory*, Ser. B,44:355-362.

[3] Dasgupta, S. 2000. Experiments with Random Projection. In *Proc.of the 16th Conference on Uncertainty in Artificial Intelligence*, 143-151.

[4] Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.