Reducing High-Dimensional Data by Principal Component Analysis vs. Random Projection for Nearest Neighbor Classification

Sampath Deegalla and Henrik Boström
Dept. of Computer and Systems Sciences,
Stockholm University and Royal Institute of Technology,
Forum 100, SE-164 40 Kista,
Sweden.
{si-sap,henke}@dsv.su.se

Abstract

The computational cost of using nearest neighbor classification often prevents the method from being applied in practice when dealing with high-dimensional data, such as images and micro arrays. One possible solution to this problem is to reduce the dimensionality of the data, ideally without loosing predictive performance. Two different dimensionality reduction methods, principle component analysis (PCA) and random projection (RP), are investigated for this purpose and compared w.r.t. the performance of the resulting nearest neighbor classifier on five image data sets and five micro array data sets. The experiment results demonstrate that PCA outperforms RP for all data sets used in this study. However, the experiments also show that PCA is more sensitive to the choice of the number of reduced dimensions. After reaching a peak, the accuracy degrades with the number of dimensions for PCA, while the accuracy for RP increases with the number of dimensions. The experiments also show that the use of PCA and RP may even outperform using the non-reduced feature set (in 9 respectively 6 cases out of 10), hence not only resulting in more efficient, but also more effective, nearest neighbor classification.

1. Introduction

With the development of technology, large volumes of high-dimensional data become rapidly available and easily accessible for the data mining community. Such data include high resolution images, text documents, gene expressions data and so on. However, high dimensional data put demands on the learning algorithm both in terms of efficiency and effectiveness. The *curse of dimensionality* is a well known phenomenon that occurs when the generation of a predictive model is mislead by an overwhelming number of features to choose between, e.g., when deciding what feature to use in a node of a decision tree [17]. Some learning methods are less sensitive to this problem since they do not rely on choosing a subset of the features, but instead base the classification on all available features. Nearest neighbor classifiers belong to this category of methods [17]. However, although increasing the number of dimensions does not typically have a detrimental effect on predictive performance, the computational cost may be prohibitively large, effectively preventing the method from being used in many cases with high-dimensional data.

In this work, we consider two methods for dimensionality reduction, principal component analysis (PCA) and random projection (RP) [4, 6, 7, 11]. We investigate which of these is most suited for being used in conjunction with nearest neighbor classification when dealing with two types of high-dimensional data: images and micro arrays.

In the next section, we provide a brief description of PCA and RP and compare them w.r.t. computational complexity. In section three, we discuss related work on these two methods. In section four, we present results from applying the methods in conjunction with nearest neighbor classification on five image data sets and five microarray data sets. Finally, we give concluding remarks and point out directions for future work.

2. Dimensionality Reduction Methods

Principal component analysis (PCA) and random projection (RP) are two dimensionality reduction methods that have been used successfully in conjunction with learning methods [4, 7]. PCA is the most well-known and popular of the above two, whereas RP is more recently gaining popularity [4, 6, 7, 11], not least by being much more efficient.

Principal component analysis (PCA)

PCA is a technique which uses a linear transformation to form a simplified data set retaining the characteristics of the original data set.

Assume that original matrix contains d dimensions and n observations and it is required to reduce the dimensionality into a k dimensional subspace then its transformation can be given by

$$Y = E^T X \tag{1}$$

Here $E_{d\times k}$ is the projection matrix which contains k eigen vectors corresponding to k highest eigen values, and where $X_{d\times n}$ is mean centered data matrix.

Random Projection (RP)

Random projection is based on matrix manipulation which uses a random matrix to project the original data set into low dimensional subspace [4, 7].

Assume that it is required to reduce the d dimensional data set into k dimensional set where number of instances are n,

$$Y = RX \tag{2}$$

Here $R_{k\times d}$ is the random matrix and $X_{d\times n}$ is the original data matrix. The idea underlying random projection originates from the Johnson-Lindenstrauss lemma (JL) [5]. It states that n points could be projected from $R^d \rightarrow R^k$ while preserving the Euclidean distance between points within an arbitrarily small factor. For the theoretical effectiveness of random projection method, see [7].

Several algorithms have been proposed to generate random projections with the same properties as JL, and the algorithms introduced by Achlioptas [1] have received significant attention [4,7]. According to Achlioptas, the elements of the random vector R can be constructed in the following way:

$$r_{ij} = \begin{cases} +1 & \text{with} \quad P_r = \frac{1}{2}; \\ -1 & \text{with} \quad P_r = \frac{1}{2}. \end{cases} \tag{3}$$

$$r_{ij} = \begin{cases} +\sqrt{3} & \text{with} \quad P_r = \frac{1}{6}; \\ 0 & \text{with} \quad P_r = \frac{2}{3}; \\ -\sqrt{3} & \text{with} \quad P_r = \frac{1}{6}. \end{cases} \tag{4}$$

An analysis of the computational complexity of random projection shows that it is very efficient compared to principal component analysis. Random projection requires only $O(dkn)$, whereas principal component analysis needs $O(d^2n) + O(d^3)$ [4].

3. Related work

Fradkin and Madigan [7] have compared PCA and RP with decision trees(C4.5), k-nearest-neighbor method with k=1 and k=5 and support vector machines for supervised learning. In their study, PCA outperformed RP, but it was also realized that there was a significant computational overhead of using PCA compared using RP.

Bingham and Mannila [4] have also compared RP with several other dimensionality reduction methods such as PCA, singular value decomposition (SVD), Latent semantic indexing (LSI) and Discrete cosine transform (DCT) for image and text data. The criteria chosen for the comparison was the amount of distortion caused by the method used on the original data and computational complexity. They also extended their experiments to determine the effects on noisy images and noiseless images. It was found that RP not sensitive to impulse noise and the amount of distortion caused by RP is quite the same as PCA. They have not considered above methods in supervised learning. However, they have pointed out the use of above methods in supervised learning with nearest neighbor.

Fern and Brodley [6] have used random projections for unsupervised learning. They have experimented with using RP for clustering of high dimensional data using multiple random projections with ensemble methods. Furthermore, they also compared their approach with single random projections and PCA for EM clustering. The use of multiple random projections based ensemble method outperforms PCA (forming better clusters) for all three data sets used in the study.

Kaski [11] used RP in the WEBSOM system for document clustering. RP was compared to PCA for reducing the dimensionality of the data in order to construct Self-Organized Maps. They conclude that their results using RP is as good as use of PCA. It was also found that level of saturation in RP is higher than that of PCA.

4. Empirical study

4.1. Data sets

Five image data sets and five micro array data sets are considered in this study, representing two types of high-dimensionality classification tasks.

The image data sets consist of two medical image data sets (IRMA [12], MIAS [13]), two object recognition data sets (COIL-100 [14], ZuBuD [9]) and a texture analysis data set (Outex - *TC_00013* [15]). The IRMA (Image Retrieval and Medical Application) data set contains radiography images of 57 classes, where the quality of the images varies significantly. The COIL-100 (Columbia university image library) data set consists of images of 100 objects, while

Table 1. Description of data.

Data set	Instances	Attributes	# of Classes
IRMA	9000	1024	57
COIL100	7200	1024	100
ZuBuD	1005	1024	201
MIAS	322	1024	7
Outex	680	1024	68
Colon Tumor	62	2000	2
Leukemia	38	7129	2
Central Nervous	60	7129	2
Srbct	63	2308	4
Lymphoma	62	4026	3

ZuBuD (Zurich Building Image Database) contains images of 201 buildings in Zurich city. MIAS (The Mammography Image Analysis Society) mini mammography database contains mammography images of 7 categories and finally Outex (University of Oulu Texture Database) image data set contains images of 68 general textures. The five micro array data sets are: Leukemia [8], Colon Tumor [3], Central Nervous [16], Srbct (small, round, blue, cell tumors) [10] and Lymphoma [2].

4.2. Experimental setup

For all image data sets, colour images have been converted into gray scale images and then resized into 32×32 pixel sized images, and where the brightness values are the only considered features. Therefore, all image data sets contain 1024 attributes. The number of instances and attributes for all data sets are shown in Table 1.

MATLAB® has been used to transform the original matrices into projected matrices using PCA, through the singular value decomposition (SVD) implementation of PCA. The Waikato Environment for Knowledge Analysis (WEKA) [17] has been used for RP (as described in 4) as well as for the nearest neighbor classifier. The accuracies were estimated using ten fold cross validation, and the results for RP is the average from 30 runs to account for its random nature.

4.3. Experimental results

The accuracies of using a nearest neighbor classifier on data reduced by PCA and RP, as well as without dimensionality reduction, are shown in Fig. 1 for various number of dimensions.

The experimental results show that reducing the dimensionality using PCA results higher accuracy for most of the data sets. In Table 2, it can be seen that only a few principal components is required for achieving the highest accuracy. However, RP typically requires a larger number of dimensions compared to PCA to obtain a high accuracy.

Table 2. Highest prediction accuracy obtained by nearest neighbor classifier with dimensionality reduction methods (no. of dimensions in parentheses).

Data set	RP		PCA		Original
IRMA	67.01	(250)	**75.30**	(40)	68.29
COIL100	98.79	(250)	98.90	(30)	**98.92**
ZuBuD	54.01	(250)	**69.46**	(20)	59.80
MIAS	44.05	(5)	**53.76**	(250)	43.17
Outex	21.04	(15)	**29.12**	(10)	19.85
Colon Tumor	80.22	(150,200)	**83.05**	(10)	77.42
Leukemia	91.32	(150)	**92.83**	(10)	89.47
Central Nervous	58.22	(150)	**66.33**	(50)	56.67
Srbct	93.23	(200)	**96.45**	(10)	87.30
Lymphoma	97.80	(250)	**99.86**	(20)	98.38

Classification accuracy using PCA typically has its peak for a small number of dimensions, after which the accuracy degrades. In contrast to this, the accuracy of RP generally increases with the number of dimensions. Hence, this shows that PCA is more sensitive to the choice of the number of reduced dimensions than RP. However, for all the data sets used in this study, the maximum accuracy obtained by using PCA is higher than the maximum accuracy obtained by using RP. This means that one can expect PCA to be more effective than RP if the number of dimensions is carefully chosen. The experiments also show that the use of PCA and RP may even outperform using the non-reduced feature set (in 9 respectively 6 cases out of 10).

The time required for performing a prediction is significantly reduced when using dimensionality reduction method as shown in Table 3. In table 3 shows the time required to test instances on training data with the change of dimensions. In summary, a significant speedup in classification time can be achieved when using PCA and RP, which often also lead to more accurate predictions.

5. Concluding remarks

We have compared using PCA and RP for reducing dimensionality of data to be used by a nearest neighbor classifier. Results on five image data sets and five micro array data sets show that PCA is more effective for severe dimensionality reduction, while RP is more suitable when keeping a high number of dimensions (although a high number is not always optimal w.r.t. accuracy). We observed that the use of PCA resulted in the highest accuracy for 9 of the 10 data sets. For several data sets, we noticed that both PCA and RP outperform using all features for classification. This shows that the use of PCA and RP, may not only lead to more efficient, but also more effective, nearest neighbor classification.

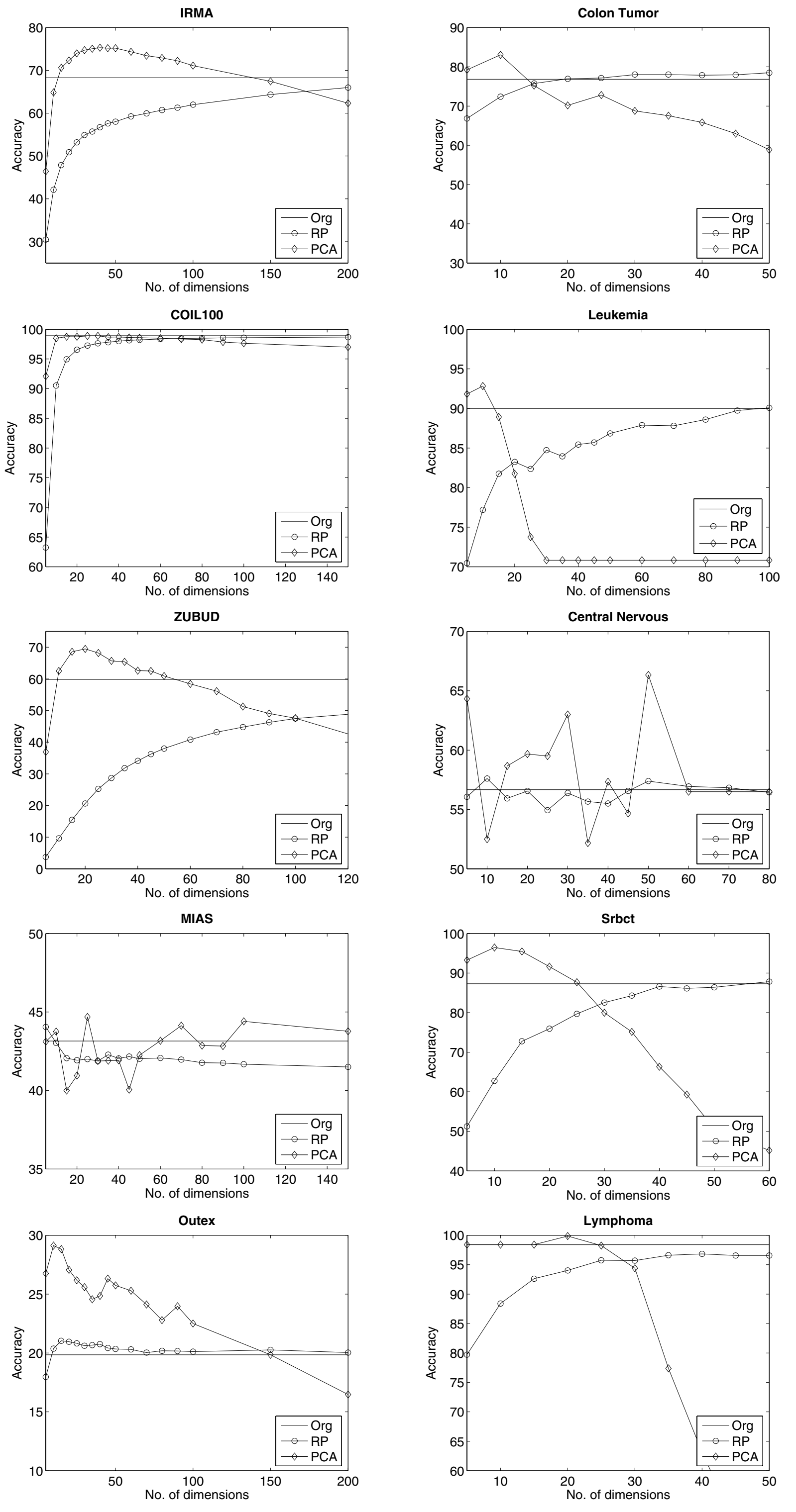

Figure 1. Comparison of the accuracies between Original, PCA and RP based attributes.

Table 3. Average time needed to test on training data (in seconds).

	IRMA	COIL-100	ZuBuD	MIAS	Outex	Colon	Leukemia	Cen. Ner.	Srbct	Lymphoma
5	71	46	0.93	0.11	0.44	0.01	0.01	0.01	0.02	0.02
10	137	87	1.76	0.20	0.81	0.02	0.01	0.01	0.01	0.01
15	207	129	2.57	0.27	1.17	0.03	0.01	0.02	0.02	0.02
20	278	172	3.39	0.37	1.56	0.02	0.02	0.02	0.02	0.02
25	344	216	4.20	0.45	1.94	0.02	0.01	0.02	0.02	0.02
30	404	258	5.01	0.53	2.31	0.02	0.01	0.03	0.03	0.03
35	478	339	5.84	0.61	2.69	0.03	0.01	0.03	0.03	0.03
40	541	344	6.65	0.70	3.08	0.03	0.01	0.03	0.04	0.04
45	609	388	7.44	0.78	3.57	0.03	0.01	0.03	0.03	0.04
50	676	433	8.25	0.87	3.86	0.04	0.01	0.03	0.04	0.04
60	809	517	9.91	1.05	4.55	0.04	0.02	0.04	0.04	0.04
70	941	617	11.57	1.23	5.44	0.05	0.02	0.05	0.05	0.05
80	1073	698	13.20	1.38	6.48	0.05	0.03	0.06	0.06	0.06
90	1206	770	14.82	1.56	6.79	0.06	0.03	0.06	0.06	0.07
100	1429	855	16.63	1.76	7.59	0.06	0.03	0.06	0.08	0.07
150	1998	1275	24.60	2.55	11.36	0.07	0.04	0.10	0.11	0.10
200	3279	1698	32.87	3.45	15.02	0.10	0.05	0.14	0.13	0.14
250	3354	2175	41.47	4.33	18.75	0.14	0.07	0.16	0.17	0.17
All	13399	8618	168.23	18.87	96.13	1.29	1.77	5.02	1.51	2.81

One direction for future work is to consider other types of high-dimensional data to gain a further understanding of the type of data for which each of the two dimensionality reduction techniques is best suited.

Acknowledgements

The authors would like to thank T.M. Lehmann, Dept. of Medical Informatics, RWTH Aachen, Germany for providing the ”10000 IRMA images of 57 categories”.

Financial support for the first author by SIDA/SAREC is greatly acknowledged.

References

[1] D. Achlioptas. Database-friendly random projections. In *ACM Symposium on the Principles of Database Systems*, pages 274–281, 2001.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. In *Proc. Natl. Acad. Sci. USA*, volume 96, pages 6745– 6750, 1999. Data set : http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

[4] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *KDD ’01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001.

[5] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical Report TR-99-006, International Computer Science Institute, Berkeley, California, USA, 1999.

[6] X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the Twentieth International Conference of Machine Learning*, 2003.

[7] D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[8] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999. Data set : http://sdmc.lit.org.sg/GEDatasets/Datasets.html.

[9] H.Shao, T. Svoboda, and L. V. Gool. Zubud - zurich building database for image based recognition. Technical report, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, 2003. Data set : http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html.

[10] J. Kahn, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[11] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN’98, International Joint Conference on Neural Networks*, volume 1, pages 413–418, Piscataway, NJ, 1998. IEEE Service Center.

[12] T. M. Lehmann, B. B. Wein, J. Dahmen, J. Bredno, F. Vogelsang, and M. Kohnen. Content-based image retrieval in medical applications: a novel multistep approach. In M. M. Yeung, B.-L. Yeo, and C. A. Bouman, editors, *Proceedings of SPIE: Storage and Retrieval for Media Databases 2000*, volume 3972, pages 312–320, 2000. Data set : http://phobos.imib.rwth-aachen.de/irma/datasets_en.php.

[13] MIAS data set : http://www.wiau.man.ac.uk/services/MIAS/-MIASmini.html.

[14] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical report, CUCS-006-96, February 1996. Data set : http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html.

[15] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, and S. Huovinen. Outex - new framework for empirical evaluation of texture analysis algorithms. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1*, 2002. Data set : http://www.outex.oulu.fi.

[16] S. L. Pomeroy, P. Tamayo, M. Gassenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, January 2002.

[17] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.