International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 8 (March 2014) PP: 46-51

Dimensionality Reduction for Classification with High-Dimensional Data

Neha Singh[1], Mr. Bharat Bhushan Agarwal[2]
[1, 2] *(Computer Science Department, Iftm University, Moradabad)*

Abstract: *High dimensional data presents a challenge for the clas-sification problem because of the diffculty in modeling the precise relationship between the large number of the class variable and feature variables. In such cases, it can* ***be*** *desirable to reduce the information for a small number of dimensions in order to improve the accuracy and effectiveness of the classification process. While data reduction has been a well studied problem for the unsupervised domain, this technique has not been explored as extensively for the supervised case. for practical use in the high dimensional case the existing techniques which try to perform dimensional- ity reduction are too slow. These techniques find global discriminants in the data. However, the data behavior often varies with data locality and different subspaces may show better discrimination in different localities. This is the more challenging task than the global discrimination problem because of the data localization issue. In this paper, I propose the PCA(Principal Component Analysis) method in order to create a reduced representation of the data for classification applications in an efficient and effective way. Because of this method, the procedure is extremely fast and scales almost linearly both with data set size and dimensionality.*
Keywords: *classification, dimensionality reduction.*

I. Introduction

The problem of dimension reduction is introduced as a way to overcome the curse of the dimensionality when dealing with vector data in high-dimensional spaces and as a modelling tool for this data. It is defined as the search to a low-dimensional manifold that embeds the high-dimensional data and A classification of dimension reduction problems is proposed.

1.1 Motivation

Consider an application in which a system processes data in the form of a collection of real-valued vectors. Suppose that the system is only effective if the dimension of each individual vector is not too high, where high depends on the particular application. The problem of dimension reduction comes when the data is in fact of a higher dimension than tolerated. For example, in this condition take the following typical cases:
1. A face recognition or classification system based on m*n greyscale images which by the row concatenation may be transformed into mn-dimensional real vectors. In practice, one could have images of m = n = 256, or 65536-dimensional vectors; if a multilayer perceptron was to be used as the classification system,then the number of weights would be exceedingly large. Therefore we need to reduce the dimension,A crude solution would be to simply scale down the images in a manageable size. More elaborate approaches exist, e.g. in [31], a 1st neural network is used to reduce the vector dimension (which actually performs a principal component analysis of the training data) by the original dimension of 63 _ 61 = 3843 to 80 components and a second one for actually performing the classification.

2. A statistical analysis for a multivariate population. Typically there will be a few variables and the analyst is interested for finding clusters or other structure of the population and/or interpreting the variables. for that aim, it is quite convenient for visualising the data, but this will not be reasonably possible for more than 3 dimensions. For example, in [73] data for the three species of eabeetles Ch. concinna, Ch. heptapotamica and Ch. heikertingeri are given and For each individual, 6 measurements were taken, including head width, front angle of the aedeagus and others. In this case, reduction to a two or three-dimensional space from the projection pursuit techniques can easily show the clustered structure of the data.

3. A more straightforward example is the estimation of a function of several variables by a finite sample. Due to the curse of the dimensionality, we may greatly reduce the sample size by reducing the number of variables, i.e. the dimension of the data.Therefore, in a number of occasions it may be useful or necessary to first reduce the dimension of the data to a manageable size, keeping as much of the original information as possible, and then feed the data of reduced-dimension into the system. Figure 1 summarises this situation and this is showing the dimension reduction as a preprocessing stage in the whole system.

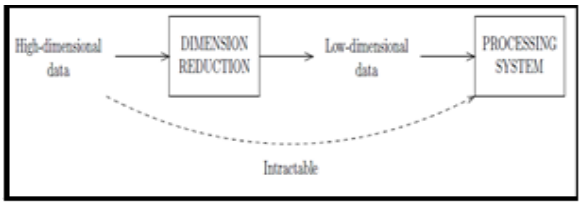

Figure 1: The dimension reduction problem.

A given processing system is only effective with vector data that is of not more than a certain dimension, so data of higher dimension must be reduced before being fed into the system.
Sometimes, a phenomenon which is in appearance high-dimensional, and thus complex, can actually be governed from the few simple variables (sometimes called \hidden causes" or \latent variables" [29, 20,21, 6, 74]). Dimension reduction may be a powerful modelling tool such phenomena and improve our understanding of them (as often the new variables will have an interpretation). For example:

4. A Genome sequences modelling. A protein is a sequence of aminoacids (of which there are 20 different ones) with residue lengths varying by the tens to tens of thousands. Proteins with the same spatial structure but oftenly with very different aminoacid sequences that are grouped together in families.
A model of protein families may give insight into the properties of particular families and may also help for identifying new members of a family or to discover new families. Probabilistic approaches for the investigation of the structure of protein families include hidden Markov models [70] and density networks [74].

5. A Speech modelling. It has been conjectured that speech recognition, undoubtedly it is an exceedingly complex process, it could be accomplished with only about 5 variables.

II. Why Is Dimension Reduction Possible?

Often, the original representation of the data will be redundant for many of the reasons:

1. Several variables will have a variation smaller than the measurement noise and thus it will be irrelevant.
2. Many of the variables will be correlated with each other, a new set of incorrelated variables should be found.

Therefore in several situations it should be possible to somehow strip of the redundant information, producing a more economic representation for the data.

2.1 The Curse of the Dimensionality and the Empty Space Phenomena

The curse of the dimensionality refers to the fact that, in the absence of simplifying assumptions, the size of the sample needed to estimate a function for several variables to a given degree of accuracy (i.e. to get a reasonably low-variance estimate) grows exponentially with the many of variables.
For example: Most density smoothers base their estimates on some local average of the neighbouring observations but in order to find enough neighbours in the spaces of high-dimensional, multivariate smoothers have to reach out farther and the locality may be lost.

A way to avoid the curse of the dimensionality is to reduce the input dimension for the function to be estimated; in unsupervised methods this is the basis for the use of local objective functions, depending on a small number of variables.

A related fact, responsible for the curse of the dimensionality, is the empty space phenomenon high-dimensional spaces are inherently sparse. For example, the probability that a point distributed uniformly in the unit 10-dimensional sphere falls at a distance of 0:9 /less from the centre is only 0:35. in multivariate density estimation this is a difficult problem, as regions of relatively very low density may contain a considerable part of the distribution, whereas regions of apparently high density may be completely devoid of observations in a moderate sample size. For example, for a one-dimensional standard normal N(0; 1), 70% of the mass is at points contained in a sphere of radius one standard deviation ; for a 10-dimensional N(0; I), that same (hyper)sphere contains only 0:02% of the mass and it has to take a radius of more than 3 standard deviations to contain 70%. Therefore, and contrarily for our intuition, in high-dimensional distributions the tails are much more important than in one-dimensional ones. Another problem that is caused by the curse of the dimensionality is that, if there are linear correlations in the data, the optimal mean integrated squared error when estimating the data density will be very large even if the sample size is arbitrarily large.

2.2 The Intrinsic Dimension of a Sample

Consider a certain phenomenon that is governed by L independent variables. In practice, this phenomenon will actually appear as having more degrees of freedom due to the inuence of a variety of factors: noise, imperfection in the measurement system, addition of irrelevant variables, etc. However,provided this inuence ,it is not too strong as to completely mask the original structure, we should be able to "filter" it out and recover the original variables or an equivalent set for them. We deifne the intrinsic dimension of a phenomenon as many of independent variables that explain satisfactorily that phenomenon.By a purely geometrical point of view, the intrinsic dimension would be the dimension L of the manifold that embeds a unknown distribution sample in D-dimensional space (D > L),satisfying certain smoothness constraints. Incidentally, we know by set theory that card

(RD) = card (R) (= N0) for any D$\in$ N, which

Means that we can map invertibly and continuously RD into R, for example using the

Diagonal Cantor construction5. In principle, this would allow to find a (nonlinear) continuous mapping from RD into RL, L < D, preserving all information. Of course, due to the finite precision this is not a practical application.

The determination of the intrinsic dimension for a distribution given a sample of this is central to the problem of dimension reduction, because knowing it would eliminate the possibility of over- or under- fitting. All the dimension reduction methods known to the author take the intrinsic dimension as a parameter that is given by the user; a trial-and-error process is necessary to obtain a satisfactory value for it Of course, this problem is itself ill-posed, because given a sample of data, it is possible for making a manifold of any dimension pass through it with negligible error given enough parameters. For example, in fig. 3 one-dimensional manifold (the dotted curve) is forced for interpolate a set of points which naturally would lie on the 2-dimensional manifold shown (the dotted rectangle). This is necessary to introduce some a priori knowledge about the degree of smoothness of manifold, perhaps in terms of a regularization term in the certain objective function.

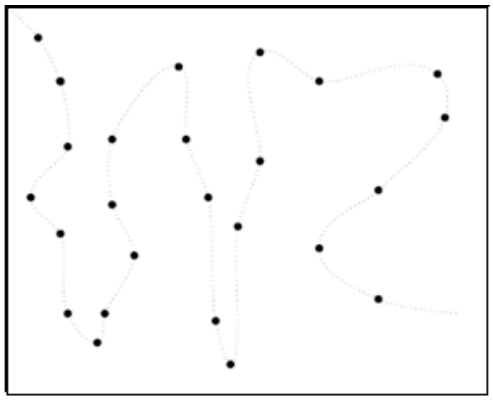

Figure 3: Curve or surface?

2.3 Views of the Problem Of Dimension Reduction

Given the basic nature of the curse of the dimensionality, this is not surprising that many different fields are affected from this. We may look at the dimension reduction problem by a number of perspectives:

1. Basically, this is nothing else but a projection, i.e. a mapping from a D-dimensional space onto an L-dimensional one, for D > L, with this associated change of coordinates.
2. In statistics, this is related to multivariate density estimation, regression and smoothing techniques.
3. By the pattern recognition standpoint, this is equivalent to feature extraction, where the feature Vector would be the reduced-dimension one.
4. In information theory, this is related to the problem of data compression and coding.
5. Many visualization techniques are actually performing some kind of dimension reduction: multidimensional scaling, Summon mapping, etc.
6. A Complexity reduction: if the complexity in time or memory of an algorithm depends on the dimension of its input data as a consequence of the curse of the dimensionality, reducing this will make the algorithm more efficient.
7. A Latent-variable models: in the latent-variable approach, we assume that a small many of hidden causes acting in combination gives rise to the apparent complexity of the data.

2.4 Classes of Dimension Reduction Problems
We attempt here a rough classification for the dimension reduction problems:

1. A Hard dimension reduction problems, in which the data have dimension ranging from hundreds to perhaps hundreds of thousands of components, and usually a drastic reduction is sought. The components are often repeated measures of a certain magnitude in different points of space or in different instants of time. In this class we would find pattern recognition and classifcation problems involving images or speech. Principal component (PCA)analysis is one of the most widespread techniques in many of the practical cases.

2. A Soft dimension reduction problems, in which the data is not too high-dimensional, and the reduction is not very drastic. Typically, the components are observed or measured values of different variables, which have the straightforward interpretation. Most statistical analyses in fields like social sciences, psychology, etc. fall in this class. Typical methods include all the usual multivariate analysis methods [75]: PCA, factor analysis, discriminant analysis, multidimensional scaling, etc.

3. Visualisation problems, in which the data doesn't normally have a very high dimension in absolute terms, but there we need to reduce it to 2, 3 or 4 at most in order to plot it. Lots of applications by many of the disciplines fall into this category. Many of the methods are used in practice, including projection pursuit,PCA, multidimensional scaling and self-organising maps and their variants, among others, and well as interactive programs that allow manual intervention(such as Xgobi [95]).

2.5 Method for Dimension Reduction
Principal Component Analysis

Principal component analysis7 (PCA) is possibly the dimension reduction technique most widely used in practice, perhaps due to its conceptual simplicity and to the fact that relatively efficient algorithms (of polynomial complexity) exist for its computation. In signal processing it is known as the Karhunen-Loeve transform.

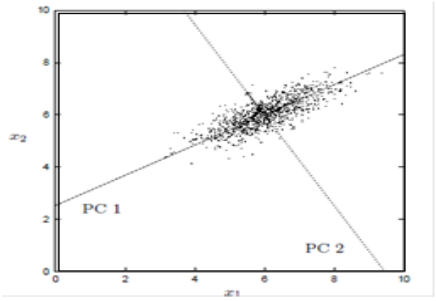

Figure 4: Bi-dimensional, normal point cloud with its principal components.

Geometrically, the hyperplane spanned by the first L principal components is the regression hyperplane that shorten the orthogonal distances to the data. In this sense, principal component analysis is a symmetric regression approach, as opposed to standard linear regression, which points one component as response variable and the rest as predictors.
The key property of PCA is that it attains the best linear map x $\in$ RD -> x* $\in$ RL in the senses of:

1. Least squared sum of the errors of the reconstructed data.
2. Maximum mutual information between the Original vectors x and their projections x*: I(x; x*) = ½ ln ((2*22/7*e)L a1……aL), where a1……aL are the first Leigenvalues of the covariance matrix.

The first principal components are often used as starting points for other algorithms, such as projection pursuit regression, principal curves, Kohonen's maps or the generalised topographic mapping, all of which are reviewed in this report.

There exist several neural network architectures capable to extract principal components;Also, when the data is clustered, it can be more convenient to apply principal component analysis locally. For example, in piecewise PCA (Kambhatla and Leen [67]), a partition of RD is defined by some form of vector quantisation of the data set and PCA applied locally in each region. This approach is fast and has comparable results to autoencoders.

A number of numerical techniques exist for finding all or the first few Eigen values and eigenvectors of a square, symmetric, semide_nite positive matrix (the covariance matrix) in O(D3): singular value decomposition, Cholesky decomposition, etc.; see [81] or [99]. When the covariance matrix, of order D * D, is too large to be explicitly computed one could use neural network techniques (section 6.2), some of which do not require more memory space other than the one needed for the data vectors and the principal components themselves. Unfortunately, these techniques (usually based on a gradient descent method) are much slower than traditional methods.The disadvantages of PCA are:

1. This is only able for finding a linear subspace and thus cannot deal properly with data lying on nonlinear manifolds.
2. One does not know how many principal components to keep, although some thumb rules are applied. For example, eliminate components whose eigenvalues are smaller than a fraction of the mean eigenvalue, or keep as many as necessary for explaining a certain fraction of the total variance [27].

III. Conclusion And Future Work

We have defined the problem of dimension reduction like a search for an economic coordinate representation of a sub manifold of a high-dimensional Euclidean space, a problem is not yet solved in a satisfactory and general way. We have then given a method of well-known technique for dimension reduction. The two major issues that are remain open:

1. To overcome the curse of the dimensionality, this demands huge sample sizes for obtaining reasonable results. Most of the techniques reviewed still suffer of this to an extent.
2. To determine the intrinsic dimension for a distribution given a sample for this. It is central to the problem of dimension reduction, because knowing this would eliminate the possibility of over- or understating.

Finally, we would like to conclude by mentioning many of further techniques that are related to dimension reduction that have not been included in this work due to lack of time. These would include the Helmholtz machine [19, 20], some variations of self-organising maps (growing neural gas [39, 40], Bayesian approaches, population codes, and curvilinear component analysis [22], among others.39

REFRENCES

[1] D. Asimov, The grand tour: A tool for viewing multidimensional data, SIAM J. Sci. Stat. Comput.,6 (1985), pp. 128{143.

[2] P. Baldi and K. Hornik, Neural networks and principal component analysis: Learning from examples without local minima, Neural Networks, 2 (1989), pp. 53{58.

[3] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, 1961.

[4] E. L. Bienenstock, L. N. Cooper, and P. W. Munro, Theory for the development of neuron selectivity: Orientation speci_city and binocular interaction in visual cortex, J. Neurosci., 2 (1982), pp. 32{48.

[5] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, Oxford, 1995.

[6] C. M. Bishop, M. Svens_en, and C. K. I. Williams, EM optimization of latent-variable density models, in Advances in Neural Information Processing Systems, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds., vol. 8, MIT Press, Cambridge, MA, 1996, pp. 465{471.

[7] GTM: A principled alternative to the self-organising map, in Advances in Neural Information Processing Systems, M. C. Mozer, M. I. Jordan, and T. Petsche, eds., vol. 9, MIT Press, Cambridge, MA, 1997.

[8] Magni_cation factors for the GTM algorithm, tech. rep., Neural Computing Research Group, Aston University, 1997.

[9] H. Bourlard and Y. Kamp, Autoassociation by the multilayer perceptrons and singular value decomposition, Biological Cybernetics, 59 (1988), pp. 291{294.

[10] L. J. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classi_cation and Regression Trees, Wadsworth, Belmont, Calif., 1984.

[11] M. _A. Carreira-Perpi~n_an, Compression neural networks and feature extraction: Application to human recognition from ear images, Master's thesis, Facultad de Inform_atica, Technical University of Madrid, Sept. 1995.

[12] B. Cheng and D. M. Titterington, Neural networks: A review from a statistical perspective, Statistical Science, 9 (1994), pp. 2{30 (with comments, pp. 31{54).

[13] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, J. Amer. Stat. Assoc., 68 (1973), pp. 361{368.

[14] D. Cook, A. Buja, and J. Cabrera, Projection pursuit indexes based on orthonormal function expansions, Journal of Computational and Graphical Statistics, 2 (1993), pp. 225{250.

[15] G. W. Cottrell, P. W. Munro, and D. Zipser, Image compression by backpropagation: A demonstration of extensional programming, in Advances in Cognitive Science, N. E. Sharkey, ed., vol. 2, Abbex, Norwood, NJ, 1988.

[16] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications, John Wiley & Sons, New York, London, Sydney, 1991.

[17] J. D. Cowan, G. Tesauro, and J. Alspector, eds., Advances in Neural Information Processing Systems, vol. 6, Morgan Kaufmann, San Mateo, 1994.

[18] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control, Signals and Sys., 2 (1989), pp. 304{314.

[19] P. Dayan and G. E. Hinton, Varieties of Helmholtz machine, Neural Networks, 9 (1996), pp. 1385{1403.

[20] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, The Helmholtz machine, Neural Computation, 7 (1995),pp.889{904}

[21] P. Dayan and R. S. Zemel, Competition and multiple cause models, Neural Computation, 7 (1995), pp. 565{579.

[22] P. Demartines and J. H_erault, Curvilinear Component Analysis: a self-organizing neural net- work for nonlinear mapping of data sets, IEEE Trans. Neural Networks, 8 (1997), pp. 148{154.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, B, 39 (1977), pp. 1{38.
[24] P. Diaconis and D. Freedman, Asymptotics of graphical projection pursuit, Annals of Statistics, 12 (1984), pp. 793{815.
[25] P. Diaconis and M. Shahshahani, On nonlinear functions of linear combinations, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 175{191.
[26] K. I. Diamantaras and S.-Y. Kung, Principal Component Neural Networks. Theory and Appli- cations, Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons, New York, London, Sydney, 1996.
[27] G. H. Dunteman, Principal Component Analysis, no. 07{069 in Sage University Paper Series on
Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills, 1989. [28] G. Eslava and F. H. C. Marriott, Some criteria for projection pursuit, Statistics and Computing, 4 (1994), pp. 13{20.
[29] B. S. Everitt, An Introduction to Latent Variable Models, Monographs on Statistics and Applied Probability, Chapman & Hall, London, New York, 1984.
[30] S. E. Fahlman and C. Lebiere, The cascade-correlation learning architecture, in Advances in Neural Information Processing Systems, D. S. Touretzky, ed., vol. 2, Morgan Kaufmann, San Mateo, 1990, pp. 524{532.
[31] M. K. Fleming and G. W. Cottrell, Categorization of faces using unsupervised feature ex-traction, in Proc. Int. J. Conf. on Neural Networks, vol. II, 1990, pp. 65{70.
[32] P. F□oldi_ak, Adaptive network for optimal linear feature extraction, in Proc. Int. J. Conf. on Neural Networks, vol. I, 1989, pp. 401{405.
[33] J. H. Friedman, A variable span smoother, Tech. Rep. 5, Stanford University, 1984.
[34] Exploratory projection pursuit, J. Amer. Stat. Assoc., 82 (1987), pp. 249{266.
[35] Multivariate adaptive regression splines, Annals of Statistics, 19 (1991), pp. 1{67 (with comments, pp. 67{141).
[36] J. H. Friedman and W. Stuetzle, Projection pursuit regression, J. Amer. Stat. Assoc., 76 (1981), pp. 817{823.
[37] J. H. Friedman, W. Stuetzle, and A. Schroeder, Projection pursuit density estimation, J. Amer. Stat. Assoc., 79 (1984), pp. 599{608.
[38] J. H. Friedman and J. W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Computers, C{23 (1974), pp. 881{889.
[39] B. Fritzke, Growing cell structures |a self-organizing network for unsupervised and supervised learning, Neural Networks, 7 (1994), pp. 1441{1460.
[40] Some competitive learning methods, draft, Institute for Neural Computation, Ruhr-Universit□at Bochum, 5 Apr. 1997.
[41] R. M. Gray, Vector quantization, IEEE ASSP Magazine, (1984), pp. 4{29.
[42] P. Hall, On polynomial-based projection indices for exploratory projection pursuit, Annals of Statistics, 17 (1989), pp. 589{605.
[43] W. J. Hardcastle, F. E. Gibbon, and K. Nicolaidis, EPG data reduction methods and their implications for studies of lingual co articulation, J. of Phonetics, 19 (1991), pp, 251-266.