International Journal of Computer Applications in Engineering Sciences
[VOL II, ISSUE II, JUNE 2012] [ISSN: 2231-4946]

A Study of Algorithms for Dimensionality Reduction

Ashoke Kr. Brahma

Dept. of Mathematics, Science College, Kokrajhar, Assam

Abstract **- It is found from studies that the sizes of real-world databases are usually very large. This is basically due to the excessive number of features or due to the huge number of instances or both. It may not give appropriate results on analysing the data considering this excessive feature space. Reduction in the data size can help the data mining tools to work efficiently. There are various algorithms developed in the aim of dimension reduction. In this paper, a study on some of the influential dimension reduction algorithms used in dimensionality reduction is carried out.**

Keywords **– Dimension Reduction, Algorithms, PCA, Clustering, High dimensional data**

I. INTRODUCTION

The term "curse of dimensionality" was first introduced by Bellman in the year 1961. It refers to the problems associated with multivariate data analysis as the dimensionality increases. There are huge mathematical challenges has to be encountered with high-dimensional datasets. One of the problems generally found with high-dimensional datasets is that, in many cases, not all the measured variables are important for understanding the underlying pattern of interest. While certain computationally expensive novel methods [26] can construct predictive models with high accuracy from high-dimensional data, it is still of interest in many applications to reduce the dimension of the original data prior to any modelling of the data.

The increasing amount of "raw" data coming from various scientific experiments have led to a persistent demand for modelling approaches which allow to extract physically interpretable information out of the data. The requirement is an automatized process for generation of low dimensional physical models based on noisy data. In this regard some interesting approaches that may provide data–based dimension reduction be a great help. This should be carefully distinguished from analytical approaches like, the Zwanzig-Mori approach, the Karhunen-Lo`ve Expansion, or averaging techniques. The latter approaches allow to reduce the dimension of a given physical model, but the problem of finding essential coordinates must be solved previously and may be data–driven as well. More information in this can be availed from [27] and [28].

In this chapter some of the influential algorithms used for dimension reduction are discussed.

II. ALGORITHMS FOR DIMENSION REDUCTION

Researchers are always trying to develop algorithms for efficient data analysis and finding interesting patterns. For high-dimensional data first they have used various statistical and mathematical techniques to remove unwanted dimensions. Later, various sophisticated algorithms or methods have been developed based on these statistical and mathematical techniques for dimension reduction i.e. clustering algorithms. Some of the dimension reduction algorithms are discussed below:

Yongming Qu et. al. describes a global principal component analysis method in [2] for dimension reduction. They assume that a virtual $n \times p$ (items x features) data matrix is distributed by blocks of rows (items), where $n > p$ and the distribution among s locations is determined by a given application. Their approach is to perform local PCA on local data without any data movement and then move and merge the local PCA results into a global PCA. The representation of local data by a few local principal components greatly reduces data transfers with minimal degradation in accuracy.

Horenko et. al. present a method for simultaneous dimension reduction and metastability analysis of high dimensional time series in [3]. The approach is based on the combination of hidden Markov models (HMMs) and principal component analysis. They derived optimal estimators for the log likelihood functional and employ the Expectation Maximization algorithm for its numerical optimization. The performance of the method is derived on a generic 102-dimensional example, apply the new HMM-PCA algorithm to a molecular dynamics simulation of 12–alanine in water and interpret the results. They exploited the fact that most high-dimensional data have lower intrinsic dimensionality thus allowing a good lower-dimensional representation. Existing methods that bring data to a central location require O(np) data transfer even if only a few principal components are needed. In the worst case, when an exact PCA is computed, their algorithm is min(O(np), O($sp2$)). It is of O(sp) data transfer complexity when intrinsic dimensionality is low or when an approximate solution is sufficient.

A generic modular PCA algorithm is introduced by Moon et. al. in [4]. In the paper they had presented a

design methodology of configuring PCA based algorithms based on empirical performance results. The heart of the method is a generic modular design of PCA. It allows to systematically very the components and measure the impact of these variations on performance. The method was tested on face recognition systems and found expected results.

The researchers proposed a statistical dimension reduction approach in [5] for classification of tumors based on microarray gene expression data. They have designed the method to address the curse of dimensionality to overcome the problem of a high dimensional gene expression space so common in such type of problems. They viewed the classification problem as a regression problem, with group membership being the response variable, and used adapted nonparametric regression methods to solve the problem. The results on a real data set show that such an approach is successful. A method proposed by [Xia et al. (2002)] used by the researchers to estimate the EDR directions. It is known as the minimum average variance estimation (MAVE) method which is easy to implement and needs no strong assumptions on the probabilistic structure of X. The method also allows to estimate consistently the dimension of the EDR space. The MAVE method may be seen as a combination of nonparametric function estimation by local polynomials and direction estimation, which is executed simultaneously with respect to the directions and the nonparametric link function. The algorithm was tested with a publicly available leukemia data set. Misclassification rates for the classifiers were estimated using leave-one-out cross-validation on the training set which is the one obtained in Dudoit et al. (2002) by dividing the dataset into a learning set and a test set comprising respectively two thirds and one third of the data.

The authors in [6] analyze dimensionality reduction in the context of multi-label classification. They simultaneously performed dimensionality reduction and multi-label classification. It is known that when the least squares loss is used in classification, this joint learning decouples into two separate components. This partially justifies the current practice of a separate application of dimensionality reduction for classification problems. When other loss functions, including the hinge loss and the squared hinge loss, are employed the resulting optimization problems are non-convex. The authors proposed a simple alternating algorithm to solve the joint learning problem. Experiments done by them showed that the alternating algorithm often converges in a few steps. One appealing feature of the joint learning formulations is that they can be extended naturally to cases where the input data for different labels may differ, overcoming the limitation of traditional dimensionality reduction algorithms. The experiments were conducted using a collection of multi-label data sets. From the experiments it was confirmed that the joint formulations are comparable to a separate dimensionality reduction and classification, while they significantly outperform classification without dimensionality reduction. Superiority of the joint formulation was shown using a data set in which the input data for different labels differ, and thus traditional dimensionality reduction algorithms are not applicable. They have also found that when the least squares loss is used in classification, the joint learning decouples into two separate components. They proved that the iterative algorithm in [6] converges in a small number of iterations.

A systematic assessment of the performance of the PLS–PH regression method is discussed in [7]. An analytical study of PLS–PH regression is not tractable because the dimension reduction method (PLS) involves complex non-linear functions of both the predictors and response variable [13]. The dimension reduction performance of the methodology is assesed based on a simulation model for gene expression data with a censored response variable. The authors have presented a comparison on the relative performances of PLS dimension reduction to dimension reduction via principal components analysis. PLS is also compared to a modified PLS (MPLS) approach. This also attempts to incorporate censoring information into the dimension reduction stage. The authors have extended a simulation model for gene expression data with a censored response variable and examined the performance of the PH regression model under three dimension reduction techniques: PLS, PCA, and a modified PLS (MPLS) technique. They have found from the study that PLS–PH and MPLS–PH outperform PCA–PH overall and perform substantially better when the total predictor variance explained (TPVE) is in the range 40%–60% and it performs slightly better than MPLS–PH and is best overall. As the TPVE increases PCA–PH improves, as expected, and all methods perform similarly when the TPVE is high (e.g. at 70%). They also have found that when the censoring rate is high (e.g. at 50%) the performance of the dimension reduction methods deteriorates, although the relative patterns of performance among the methods are similar and MPLS–PH appears to be most affected by the high censoring, overestimating survival time at high TPVE (70%). From the study it was found that PLS–PH method works well, despite ignoring information on censoring at the dimension reduction stage.

The problem of dimensionality reduction is formulated as semi-parametric estimation of the low dimensional signal in [8], treating the signal distribution as unconstrained nuisance and the noise distribution as

constrained nuisance. The authors have presented an estimator which is appropriate when the conditional means E[y|u] lie in a low-dimensional *linear* space, and a maximum-likelihood estimator for additive Gaussian mixture noise.

It is based on the fact that, the maximum-likelihood low-rank estimation cannot be taken for granted, and demonstrates that it might not be consistent even for known noise models. The approach employed by them can also be used to investigate the consistency of ML estimators with non-additive noise models. Of particular interest are distributions yi|xi that form exponential families where xi are the *natural* parameters. When the *mean* parameters form a low-rank linear subspace, the variance-ignoring estimator is applicable, but when the natural parameters form a linear subspace, the means are in general curved, and there is no unbiased estimator for the natural parameters. According to the initial investigations the ML estimator for a Bernoulli (logistic) conditional distribution is not consistent. The problem of finding a consistent estimator for the linear-subspace of natural parameters when yi|xi forms an exponential family remains open.

An approach, capable of discovering the nonlinear degrees of freedom is described in [9]. It can solve dimensionality reduction problems using easily measured local metric information to learn the underlying global geometry of a data set. It can easily solve the problems that underlie complex natural observations, such as human handwriting or images of a face under different viewing conditions. This algorithm efficiently computes a globally optimal solution, and, for an important class of data manifolds, is guaranteed to converge asymptotically to the true structure. In that paper performance of Isomap is also presented on data sets chosen for their visually compelling structures, but the technique may be applied wherever nonlinear geometry complicates the use of PCA or MDS. Isomap complements, and may be combined with, linear extensions of PCA based on higher order statistics, such as independent component analysis (21, 22). It may also lead to a better understanding of how the brain comes to represent the dynamic appearance of objects, where psychophysical studies of apparent motion (23, 44) suggest a central role for geodesic transformations on nonlinear manifolds (25).

A method termed as locally linear embedding (LLE) is introduced in [10]. It deals with discovering compact representations of high-dimensional data. It is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs. Unlike clustering methods for local dimensionality reduction, LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. By exploiting the local symmetries of linear reconstructions, LLE is able to learn the global structure of nonlinear manifolds, such as those generated by images of faces or documents of text.

A method for clustering data in a high dimensional space based on a hypergraph model is proposed in [11]. It is based on the hypergraph partitioning algorithm HMETIS. Hypergraph based clustering holds great promise for clustering data in large dimensional spaces- According to the researchers, traditional clustering schemes such as Autoclass and Kmeans cannot be directly used in such large dimensionality data sets as they tend to produce extremely poor results. These methods perform much better when the dimensionality of the data is reduced using methods such as Principal Component Analysis, Factor Analysis etc. But experiments showed that hypergraphbased scheme produces clusters that are at least as good or better than those produced by AutoClass or Kmeans algorithm on the reduced dimensionality data sets. One of the major advantages of this scheme over traditional clustering schemes is that it does not require dimensionality reduction as it uses the hypergraph model to represent relations among the data items. The proposed model allows to effectively represent important relations among items in a sparse data structure on which computationally ecient partitioning algorithms can be used to find clusters of related items. The sparsity of the hypergraph can be controlled by using an appropriate support threshold. An additional advantage of the scheme is its ability to control the quality of clusters according to the requirements of users and domains. The amount of relationship captured in the hypergraph model can be adjusted using different levels of minimum support used in Apriori algorithm. Higher support gives better quality clusters containing smaller number of data items whereas lower support results in clustering of larger number of items in poorer quality clusters. The algorithm is also capable of correct determination of the quality of the clusters by looking at the internal connectivity of the nodes in each cluster.

A projected data stream clustering method for high dimensional data termed as HPStream is proposed in [12]. The method incorporates a fading cluster structure, and the projection based clustering methodology. It is able to find projected clusters in particular subsets of the dimensions by maintaining condensed representations of the clusters over time. The method can be incrementally updatable and is also highly scalable on both the number of dimensions and the size of the data streams. The algorithm provides better quality clusters due to using of reduced number of dimensions. The projected clustering can be treated as as a preprocessing step.

A method for extracting simple descriptions of high dimensional data sets in the form of simplicial complexes is described in [13]. The method is termed as Mapper. It is based on the idea of partial clustering of the data guided by a set of functions defined on the data. It is not dependent on any particular clustering algorithm. It can be used with any clustering algorithm. The method is actually a technique for qualitative analysis, simplification and visualization of high dimensional data sets. It also performs qualitative analysis of functions on the data sets. Generally, it is seen that real world data sets are massive in size and it is not possible to visualize and discern structure even in low dimensional projections.

Another clustering algorithm for high-dimensional datasets is proposed in [14]. The key idea involves using a cheap, approximate distance measure to efficiently divide the data into overlapping subsets. These are called as canopies. Canopies can be applied to many domains and used with a variety of clustering approaches. After that, clustering is performed by measuring exact distances only between points that occur in a common canopy. The reduction in computational cost is achieved without any loss in clustering accuracy. This technique can be incorporated with Greedy Agglomerative Clustering, K-means and Expectation-Maximization etc. The authors have proved that the canopy approach reduces computation time over a traditional clustering approach by more than an order of magnitude and decreases error in comparison to a previously used algorithm by 25%. The cheap measures can be binning, comparison of a few attributes of a complex record, or finding similarity using an inverted index.

A simultaneous dimension reduction and metastability analysis of high dimensional time seriess is proposed in [15]. The approach is based on the combination of hidden Markov models (HMMs) and local principal component analysis. Incorporation of the local PCA analysis helps to map the clustering problem into low dimensional space. They have derived optimal estimators for the log-likelihood functional and employed the Expectation Maximization algorithm for its numerical optimization. The method is useful in clustering of time series data.

The paper [16] presents the study of a sparse kernel-based method for non-linear feature extraction in the context of remote sensing classification and regression problems. The method demonstrates good capabilities in terms of expressive power of the extracted features and scalability.

All previous methods assume that there exists a linear relation between the latent variables of X and of Y. However, this might not necessarily hold, and thus non-linear versions have become necessary to solve this problem. In this context, kernel methods are a promising approach, as they constitute an excellent framework to formulate non-linear versions from linear algorithms [17], which has been demonstrated to be very useful in different application domains [18]. Some recent developments based on kernel methods have been done to obtain non-linear PLS-based algorithms from linear ones while still solving only linear equations [19], [17].

This work studied the applicability of the rKOPLS method for feature extraction and dimensionality reduction in hyperspectral imaging, both for classification and regression problems. Unlike KPLS, the rKOPLS makes the data in the feature space orthonormal. In addition, sparsity is imposed so the algorithm can efficiently deal with high dimensional input samples, such as those encountered in hyperspectral image processing problems, and scales well with the number of training samples. The sparse approximation used by rKOPLS is specially convenient in this context, given that otherwise a huge kernel matrix should be stored and processed. We have observed that the method produces similar results to SVM classifier and regression machines but with much lower computational cost and memory requirements. Next steps will consider extensive comparison with other PLS-based algorithms in large-scale remote sensing classification and regression scenarios.

The researchers present a new indexing method for high dimensional data in [20]. This method is based on the surface of dimensionality as well as recursive partitioning space. It is seen that some index structures like R-tree are effected by the “curse of dimensionality” problems. To get rid of that, Pyramid-tree type of index structures, were proposed to break the curse of dimensionality. But it is found that, for high dimensional data, the number of pyramids is often insufficient to discriminate data points when the number of dimensions is high. Its effectiveness degrades dramatically with the increase of dimensionality. Pyramid tree technology is a special case of the proposed method. The technique works as follows:

To break the curse of dimensionality, high dimensional data points are transformed to 1-dimensional values. Therefore, classical index structures such as the B+-tree can be adapted. By partitioning the space recursively, the approach overcomes the restriction of 2d pyramids in the Pyramid-tree. More pyramids are partitioned and the selection of key is improved. In future work, we will estimate the optimal number of partitions required to construct an index structure considering data distribution.

Surface based index technique filters out non-related data points for a similarity query. Smaller candidate set is desirable. It is because that we have to consume I/O and CPU costs to refine every candidate to

get exact answers. If the data space (or hypercube) is partitioned more times, the index keys are more accuracy and the candidate set becomes smaller.

III. CONCLUSION

In this paper some of the algorithms used for dimension reduction are discussed. These algorithms use various techniques from statistical techniques to data mining techniques.

REFERENCES

[1] Shlens Jonathon, A Tutorial on Principal Component Analysis Center for Neural Science, New York University New York City, NY 10003-6603 and Systems Neurobiology Laboratory, Salk Insitute for Biological Studies

[2] Y. M. Qu, G. Ostrouchov, N. Samatova, and A. Geist, Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets, Proceedings to the Second SIAM International Conference on Data Mining, April 2002.

[3] Horenko, J. Schmidt-Ehrenberg, and Ch. Schuette. Set-oriented dimension reduction: Localizing principal component analysis via hidden markov models. In LNBI: Proceedings of the 2nd International Symposium on Computational Life Science, volume 4216, pages 74{85, 2006.

[4] Hyeonjoon Moon, P. Jonathan Philips, Computational and Performance Aspects of PCA-Based Face Recognition Systems, Perceptron, 2001, Vol. 30, pp 303-321.

[5] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the ”Mathematical Challenges of the 21st Century” con-ference of The American Math. Society, Los Angeles, August 6-11. http://www.stat.stanford.edu/~donoho/Lectures/AMS2000/AMS 2000.html, 2000.

[6] Antoniadis, S. Lambert-Lacroix and F. Leblanc, Effective Dimension Reduction Methods for Tumor Classification using Gene Expression Data, Laboratoire IMAG-LMC, University Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France.

[7] Shuiwang Ji, Jieping Ye, Linear Dimensionality Reduction for Multi-label Classification, Arizona State University

[8] Danh V. Nguyen, Partial least squares dimension reduction for microarray gene expression data with a censored response, Mathematical Biosciences 193 (2005) 119–137 , Division of Biostatistics, Public Health Sciences, School of Medicine, University of California, One Shields Avenue, Davis, CA 956168638, USA, www.elsevier.com/locate/mbs

[9] Nathan Srebro Tommi Jaakkola, Linear Dependent Dimensionality Reduction, Department of Electrical Engineering and Computer Science Massachusetts Institute of Technology Cambridge, MA 02139

[10] Joshua B. Tenenbaum, Vin de Silva, John C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction

[11] [10] Sam T. Roweis, and Lawrence K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding

[12] Eui Hong Sam, Han George Karypis, Vipin Kumar Bamshad Mobasher , Hypergraph Based Clustering in High Dimensional Data Sets A Summary of Results , Department of Computer Science and Engineering ArmyHPC Research Center, University of Minnesota

[13] Charu C. Aggarwal, Jiawei Han, Jianyong Wangy Philip S. Yu, A Framework for Projected Clustering of High Dimensional Data Streams, T. J. Watson Resch. Ctr. Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004

[14] Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Eurographics Symposium on Point-Based Graphics (2007), M. Botsch, R. Pajarola (Editors)

[15] Andrew McCallumzy, Kamal Nigamy, Lyle H. Ungar , Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching, WhizBang! Labs – Research, School of Computer Science

[16] Illia Horenko 1 , Johannes Schmidt -Ehrenberg2 and Christof Schmidte, Set-Oriented Dimension Reduction : Localizing Principal Component Analysis via Hidden Markov Models, Freie Universit ̈t Berlin, Department of Mathematics and Informatics, Arnimallee 6, D-14195 Berlin, Germany

[17] Jerónimo Arenas-García and Gustavo Camps-Valls, Feature Extraction from Remote Sensing Data using Kernel Orthonormalized PLS Dept. Signal Theory and Communications. Universidad Carlos III de Madrid. Spain.Enginyeria Electrònica. Universitat de València. C/ Dr. Moliner, 50. 46100. Burjassot, València. Spain.

[18] R. Rosipal and L. J. Trejo, “Kernel partial least squares regression in reproducing kernel Hilbert space,” Journal of Machine Learning Research, vol. 2, pp. 97–123, 2001.

[19] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[20] G. Camps-Valls, J. L. Rojo-Álvarez, and M. Martínez-Ramón, Eds., Kernel Methods in Bioengineering, Signal and Image Processing, Idea Group Publishing, Hershey, PA (USA), Jan 2007.

[21] Jiyuan An, Yi-Ping Phoebe Chen, Qinying Xu and Xiaofang Zhou, A New Indexing Method for High Dimensional Dataset, Deakin University, Australia, of Tsukuba, Japan, University of Queensland, Australia

[22] P. Comon, Signal Proc. 36, 287 (1994).
J. Bell, T. J. Sejnowski, Neural Comp. 7, 1129 (1995).

[23] R. N. Shepard, S. A. Judd, Science 191, 952 (1976).

[24] [24]M. Shiffrar, J. J. Freyd, Psychol. Science 1, 257 (1990).

[25] R. N. Shepard, Psychon. Bull. Rev. 1, 2 (1994).

[26] L. Breiman. Random forests. Technical report, Department of Statistics, University of California, 2001.

[27] P. Holmes, J.L. Lumley, and G. Berkooz. Turbulence, Coherent Structures, Dynamical Systems and Symmetry. Cambridge University Press, 1996.

[28] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: Modelproblems and algorithms. Nonlinearity, 17:R55–R127, 2004.