© 2018 IJSRST | Volume 4 | Issue 8 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Comparative Analysis of Dimensionality Reduction Techniques for Machine Learning

Santhosh Voruganti*[1], Karnati Ramyakrishna[2], Srilok Bodla[3], E. Umakanth[4]

*[1] Department of IT, Assistant Professor, CBIT, Hyderabad, Telangana, India

[2] Department of MCA, Osmania University, Hyderabad, Telangana, India

[3]Department of IT, CBIT, Hyderabad, Telangana, India

[4]Department of IT, CBIT, Hyderabad, Telangana, India

ABSTRACT

Dimensionality reduction as a pre-processing step to machine learning is effective in removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection and feature extraction methods with respect to efficiency and effectiveness. In the field of machine learning and pattern recognition, dimensionality reduction is important area, where many approaches have been proposed. Aim of this paper is to reduce the dimensionality of the dataset without the loss of any information from the datasets. We have implemented three dimensionality reduction algorithms.So this three algorithms are performed on two datasets, Iris and Wines datasets and the results are analyzed.

Keywords : PCA, LDA, KPCA.

I. INTRODUCTION

Machine learning aims to build computer programs that automatically improve with experiences. In statistical language, it is simply learning from data that we generate in our day to day life. Machine learning is related to diverse disciplines as it is all about automating the process of problem solving to a larger extent. It is usually studied as a part of artificial intelligence thus relating it to computer science. As already stated it deals with data that we generate thus relating it to statistics and mathematics domain. After everything, problems that it tries to solve may have origin in any discipline like biology (in DNA analysis), medicine (medical diagnosis), and commercial purposes (like product recommendations, stock trading, and credit card fraud detection). Machine learning grew out from pattern recognition. However it has progressed dramatically over the past two decades. In AI systems, it has emerged as a popular method to develop software's for various fields like computer vision, speech recognition, natural language processing etc.

In pattern recognition, data mining, and other kinds of data analysis applications, we often face high dimensional data. For example, in face recognition, the size of a training image patch is usually larger than 60 ×60 ,which corresponds to a vector with more than 3600 dimensions. Feature extraction based on the domain knowledge can be performed to explore more important information from the face patch and may result in a lower dimensional vector, usually the remaining dimensionality is still too high for learning. And for training data without explicit background or meaning, where domain knowledge cannot directly be performed, data-driven techniques for reducing the dimensionality of features are of high demand.

In recent face-related research topics, especially for face recognition and facial age estimation, dimensionality reduction (DR) plays a tremendous important role not only because features of these two topics are hard to define and usually of really high dimensionality, but also because they are both multi-class problems. For face recognition, there may be more than a hundred of people in the database and each person has no more than a dozen of images, which results in a multi-class classification problem with limited training set. The feature dimensionality does affect the VC dimension, and under the condition of limited training data, the feature dimensionality should also be carefully considered in order to maintain generalization performance. For facial age estimation, each possible age could be seen as a class, so usually the problem is of over 60 classes, which also requires limited feature dimensionality to avoid over-fitting. Outstanding face recognition techniques such as Eigen faces, fisher faces, Laplacian faces, and face recognition based on independent component analysis (ICA) exploit different kinds of dimensionality reduction methods to directly reduce the dimensionality of face patches. DR also widely used in facial age estimation.

II. Related Work

A. DIMENSIONALITY REDUCTION

Data mining refers to the task of analysing large amount of data with intend of finding hidden patterns and trends that are not immediately apparent from summarized data. Data mining and knowledge extraction from raw data is becoming more and more important and useful as the amount and complexity of data is rapidly increasing. Data mining commonly involves four classes of tasks: Classification - arranges the data into predefined groups, Clustering - is similar to classification but the groups are not predefined, so the algorithm will try to group similar items together, Regression - attempts to find a function which models the data with the least error and Association rule learning - searches for relationships between variables. Data has become highly available now-a-days and consists of complex structures and high dimensions. In order to achieve accuracy in classification of such data, we require identifying and removing irrelevant and insignificant dimensions. The process of reducing dimensions is referred as Dimensionality Reduction. It is a crucial pre -processing step in Data Mining to improve computational efficiency and accuracy. Dimensionality reduction provides benefits such as improved dataset classification accuracy, increased computational efficiency and better visualization of dimensions.

B. What Is Dimensionality Reduction

The problem of (nonlinear) dimensionality reduction can be defined as follows. Assume we have a dataset represented in a $n \times D$ matrix X consisting of n data vectors xi ($i \in \{1, 2, \ldots, n\}$) with dimensionality D. Assume further that this dataset has intrinsic dimensionality d (where $d < D$, and often D). Here, in mathematical terms, intrinsic dimensionality means that the points in dataset X are lying on or near a manifold with dimensionality d that is embedded in the D-dimensional space. Note that we make no assumptions on the structure of this manifold: the manifold may be non-Riemannian because of discontinuities (i.e., the manifold may consist of a number of disconnected sub manifolds).

Dimensionality reduction techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining the geometry of the data as much as possible. In general, neither the geometry of the data manifold, nor the intrinsic dimensionality d of the dataset X is known. Therefore, dimensionality reduction is an ill posed problem that can only be solved by assuming certain properties of the data (such as its intrinsic dimensionality). Throughout the paper, we denote a high-dimensional data point by xi, where xi is the ith row of the D-dimensional data matrix X. The low dimensional counterpart of xi is denoted by yi, where yi is the ith

row of the d-dimensional data matrix Y.Figure 1 shows a taxonomy of techniques for dimensionality reduction. We subdivide techniques for dimensionality reduction into convex and does not contain any local optima non-convex techniques. Convex techniques optimize an objective function that, whereas non-convex techniques optimize objective functions that do contain local optima.

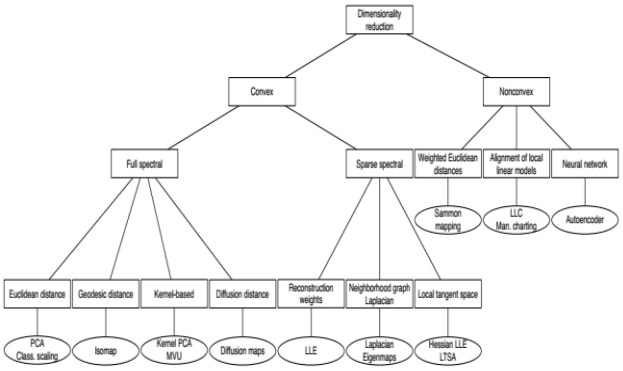


Figure 1: Taxonomy of techniques for dimensionality reduction

C. **Principal Component Analysis**:

It is a Feature Extraction technique that is used to analyse statistical data by transforming the starting set of variables into various set of linear combinations which are known as the principal components (PC), and these components have some specific properties with regard to variances. This make the dimensionality of the system more concentrated and at the same time, variable connections information is also maintained. Calculations are made on the data set by analysing eigenvalue and its eigenvectors, covariance matrix arranged systematically in descending order. This technique produces the maximum feasibility arbitrary solutions in the high-dimensional space. We can view it as data visualization method because here, the high dimensional data sets can be reduced to two dimensional or three dimensional data sets that can be easily plotted using graphs or charts.

D. Linear Discriminant Analysis:

Linear Discriminant Analysis can be used to perform supervised dimensionality reduction, by projecting the input data to a linear subspace consisting of the directions which maximize the separation between classes. The dimension of the output is necessarily less than the number of classes, so this is a in general a rather strong dimensionality reduction, and only makes senses in a multiclass setting.

LDA can be derived from simple probabilistic models which model the class conditional distribution of the data for each class Predictions can then be obtained by using Bayes rule.

E. Kernel PCA

Standard PCA only allows linear dimensionality reduction. However, if the data has more complicated structures which cannot be well represented in a linear subspace, standard PCA will not be very helpful. Fortunately, kernel PCA allows us to generalize standard PCA to nonlinear dimensionality reduction.

Assume we have a nonlinear transformation $\varphi(x)$ from the original D-dimensional feature space to an Multidimensional feature space, where usually MD. Then each data point xi is projected to a point $\varphi(xi)$. We can perform standard PCA in the new feature space, but this can be extremely costly and inefficient. Fortunately, we can use kernel methods to simplify the computation.

III. RESULTS AND DISCUSSION

In this paper we have used two datasets:

A.Iris Dataset:

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same

pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

B. **Wines Dataset:**

The data was used with many others for comparing various classifiers. The classes are separable, though only RDA has achieved 100% correct classification.(RDA : 100%, QDA 99.4%, LDA 98.9%, 1NN 96.1% (z-transformed data))(All results using the leave-one-out technique) In a classification context, this is a well posed problem with "well behaved" class structures. A good data set for first testing of a new classifier, but not very challenging.

Number of Instances:

class 1 59

class 2 71

class 3 48

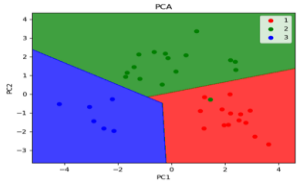
Class Distribution: number of instances per class

class 1 59

class 2 71

class 3 4 39

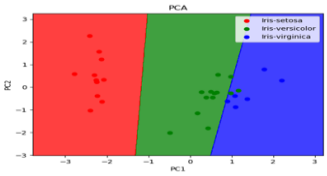
C. **PCA using Wines Dataset**




```
[[14,  0,  0],
 [ 1, 15,  0],
 [ 0,  0,  6]],
```

75 percent of dataset is used for training and 25 percent data is used for testing.

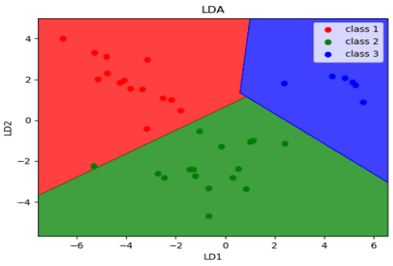
D. **PCA Using Iris Dataset:**




```
([[11,  0,  0],
 [ 0, 11,  2],
 [ 0,  0,  6]],
```

75 percent of dataset is used for training and 25 percent data is used for testing.

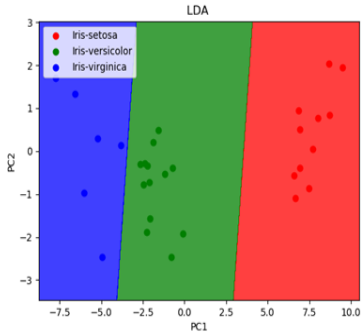
E. **LDA Using Wines dataset:**




```
([[14,  0,  0],
 [ 1, 15,  0],
 [ 0,  0,  6]],
```

75 percent of dataset is used for training and 25 percent data is used for testing. By this method efficiency is not more compared to KPCA.

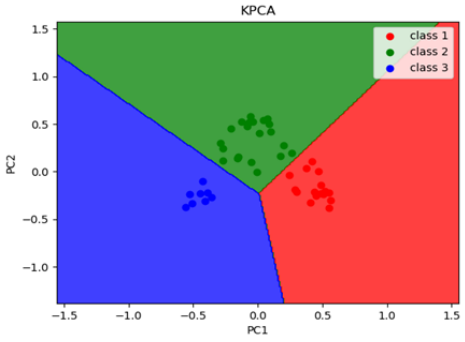
F. **LDA Using Iris Dataset**




```
[[11,  0,  0],
 [ 0, 13,  0],
 [ 0,  0,  6]],
```

75 percent of dataset is used for training and 25 percent data is used for testing. By this method efficiency is not more compared to PCA and KPCA.

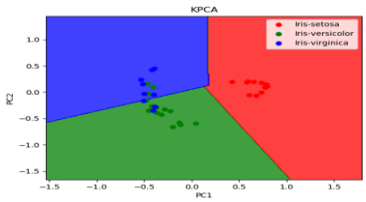
G. **KPCA Using Wines dataset:**




```
([[16,  0,  0],
 [ 0, 21,  0],
 [ 0,  0,  8]],
```

75 percent of dataset is used for training and 25 percent data is used for testing. By this method efficiency is more compared to LDA and PCA.

H.KPCA Using Iris Dataset




```
[[13,  0,  0],
 [ 0, 13,  3],
 [ 0,  3,  6]],
```

75 percent of dataset is used for training and 25 percent data is used for testing. By this method efficiency is not more compared to LDA and PCA.

IV.CONCLUSION

The primary objective of this paper is to compare various schemes that are being used to reduce the dimensionality of high dimensional datasets in order to improve accuracy and time complexity of machine learning algorithms such as classification and clustering.

The paper presented a review and comparative Analysis of techniques for dimensionality reduction. From the results obtained, we may conclude that nonlinear techniques for dimensionality reduction are, de-spite their large variance, often not capable of outperforming traditional linear techniques such as PCA. In the future, we foresee the development of new nonlinear techniques for dimensionality reduction that do not suffer from the presence of trivial optimal solutions, may be based on non-convex objective functions, and do not rely on neighbourhood graphs to model the local structure of the data manifold. The other important concern in the development of novel techniques for dimensionality reduction is their optimization, which should be computationally and numerically feasible in practice.

V.REFERENCES

1. Tom M. Mitchell, "Machine Learning ",McGraw Hill, 1997
2. Stephen Marsland, "Machine Learning - An Algorithmic Perspective ", CRC Press, 2009. Margaret H Dunham, "Data Mining", Pearson Edition, 2003.
3. GalitShmueli, Nitin R Patel, Peter C Bruce, "Data Mining, 2007 for Business Intelligence", Wiley India Edition.
4. Rajjall Shinghal, "Pattern Recognition ",Oxford University Press, 2006.
5. Ashish Kumar and Avinash Paul the authors of the book Mastering Text Mining with R,
6. Nonlinear Dimensionality ReductionAuthors: Lee, John A., Verleysen, MichelMathematical Methodologies in Pattern Recognition and MachiLearningEditors: Latorre Carmona, Pedro, S-nchez, J. Salvador, Fred - 502 854.
7. Understanding Machine Learning: From Theory to Algorithms Textbook by Shai Ben-David and ShaiShalev-Shwartz.
8. Foundations of Machine Learning Textbook by Afshin Rostamizadeh, Ameet Talwalkar, and Mehryar Mohri.K-glDonald, C. WunschAndrei, Y. ZinovyevChristopher.
9. A survey of dimension reduction techniques Imola K. FodorCenter for Applied ScientficComputing, Lawrence Livermore National Laboratory P.O. Box 808, L-560, Livermore, CA 94551.
10. IEEEI.T Jolliffe, Principal Component Analysis, Springer,second edition.
11. Chao Shi and Chen Lihui, 2005. Feature dimension reduction for microarray data analysis using locally linear embedding, 3rdAsia Pacific Bioinformatics Conference, pp. 211-217.
12. Dimensionality Reduction for Data Mining-Techniques, Applications and Trends- Jieping Ye, HuanLiuArizona State University.

13. Principal Component Analysis With Complex Kernels Athanasios Papaioannou, Student Member, IEEE, Stefanos Zafeiriou, Member, IEEE
14. A review of feature selection methods with Applications A. Jovic*, K. Brkic* and N. Bogunovic*
15. Nonlinear Multimode Industrial Process Fault Detection Using Modified Kernel Principal Component Analysis XIAOGANG DENG , NA ZHONG, AND LEI WANG College of Information and Control Engineering, China University of Petroleum, Qingdao 266580, China