[ISSN: 2231-4946]

Mathematical & Statistical Techniques for Dimension Reduction of High Dimensional Data: A Selective Survey

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. One such type of techniques is reducing the size by decreasing the number of features for easy and efficient analysis of data. There are different ways for achieving this goal, such as determine the relative importance of the features and then select a subset of important features which can be achieved in various ways, for example, by feature selection or feature extraction. In this paper, a study on the existing statistical and mathematical techniques used in dimensionality reduction is carried out. It also discusses some influential algorithms in this field**.

***Keywords*- Data, Dimension Reduction, Feature Selection, PCA, Factor Analysis**

I. INTRODUCTION

Due to improvements in data collection techniques and storage capabilities in the recent times information overload has been occurred in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis. Such datasets, in contrast with smaller, more traditional datasets that have been studied extensively in the past, present new challenges in data analysis. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation. The dimension of the data is the number of variables that are measured on each observation.

A conventional database schema may be composed of many different attributes. The problem here is that, all the attributes may not be needed to solve a given data mining problem. In fact, the use of some attributes may interfere with the correct completion of a data mining task. The use of other attributes may simply increase the overall complexity and decrease the efficiency of an algorithm. This problem is sometimes referred to as dimensionality curse, meaning that there are many attributes involved and it is difficult to determine which ones should be used. One solution to this high dimensionality problem is to reduce the number of attributes, which is known as dimensionality reduction.

A. The Problem of Traditional Clustering Algorithms in High Dimensional Data

Searching in high-dimensional spaces is time-consuming. Performing point and range queries in high dimensions is considerably easier, from the point of computational complexity, than performing similarity queries because point and range queries do not involve the computation of distance. Most clustering methods are designed for clustering low-dimensional data and face problems when the dimensionality of the data grows really high. This is because in a particular dataset usually only a small number of dimensions are relevant to the problem of interest. Therefore, data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered. Moreover, when dimensionality increases, data usually become increasingly sparse because the data points are likely located in different dimensional subspaces. When the data become really sparse, data points located at different dimensions can be considered as all equally distanced, and the distance measure, which is essential for cluster analysis, becomes meaningless.

B. Dimensionality Reduction

Dimension reduction is an essential step in high-dimensional data analysis. It extracts a small number of features by removing irrelevant, redundant, and noisy information. It is the mapping of data to a lower dimensional space such that uninformative variance in the data is discarded, or such that a subspace in which the data lives is detected. Dimension reduction is a crucial step for the analysis of high-dimensional data. Classical dimensionality reduction techniques include unsupervised algorithms such as principal component analysis [1] and supervised algorithms such as linear discriminant analysis [2], canonical correlation analysis [3], etc. These algorithms are commonly applied as a separate data preprocessing step before classification

algorithms, and they have been applied successfully too many real-world problems.
apart from its intrinsic usefulness, PCA is interesting because it serves as a starting point for many modern algorithms, some of which (kernel PCA, probabilistic PCA, and oriented PCA) are also described here.

II. MATHEMATICAL & STATISTICAL TECHNIQUES

A. Principal Component Analysis (PCA)

PCA [3] is a standard tool in modern data analysis. It is used in diverse fields from neuroscience to computer graphics. It is probably the most widely-used and well-known of the "standard" multivariate methods. It was invented by Pearson and Hotelling. PCA was first applied in ecology by Goodall in the year 1954. But he used it under the name "factor analysis". It is a simple, non-parametric method for extracting relevant information from confusing data sets. PCA is a mathematical procedure that transforms a number of correlated variables into a number of uncorrelated variables called principal components. The number of uncorrelated values are generally smaller. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. PCA can provide a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it.

PCA takes a data matrix of *n* objects by *p* variables, which may be correlated, and summarizes it by uncorrelated axes. It then tries to reduce it to n objects by k<=p variables by searching k n-dimensional orthogonal vectors that can best be used to represent the data. These are linear combinations of the original *p* variables where the first *k* components display as much as possible of the variation among objects.

The goal of principal component analysis [6] is to identify the most meaningful basis to re-express a data set. The hope is that this new basis will filter out the noise and reveal hidden structure. In the example of the spring, the explicit goal of PCA is to determine: "the dynamics are along the x-axis." In other words, the goal of PCA is to determine that x, i.e. the unit basis vector along the x-axis, is the important dimension.

B. Independent Component Analysis

ICA [6] is a higher-order method that seeks linear projections, not necessarily orthogonal to each other, that are as nearly statistically independent as possible. It is a computational method for separating a multivariate signal into additive subcomponents supposing the mutual statistical independence of the non-Gaussian source signals [7]. Statistical independence is a much stronger condition than absence of correlation. The non-correlation only involves the second-order statistics, but statistical independence depends on all the higher-order statistics. ICA can be applied to many different problems, including exploratory data analysis, blind source separation, feature extraction etc. In the feature extraction context, the columns of the matrix A represent features in the data, and the components s_i give the coefficient of the i^{th} feature in the data.

To visualise the concept of ICA a nice exmple stated in [4] can be taken. It is known as cocktail-party problem.

Let, in a room, where two people are speaking simultaneously. Two microphones are there in the room, which are placed in different locations. Those microphones can give two recorded time signals of the persons present out there. These can be denoted by x1 (t) and x2 (t), where x1 and x2 the amplitudes, and t the time index. Each of these recorded signals is a weighted sum of the speech signals emitted by the two speakers. These weighted sums can be denoted by s1 (t) and s2 (t). Finally, these can be expressed as a linear equation as shown below:

$$x1(t) = a11\ s1 + a12 s2$$
$$x2(t) = a21\ s1 + a22 s2$$

where a11 , a12 , a21 , and a22 are some parameters that depend on the distances of the microphones from the speakers. It would be very useful if you could now estimate the two original speech signals s1 (t) and s2 (t), using only the recorded signals x1 (t) and x2 (t).

Let the data is represented by a random vector $a = (a_1, a_2, \ldots a_p)$. A random vector [5] is a list of mathematical variables each of whose values is unknown, either because the value has not yet occurred or because there is imperfect knowledge of its value. The individual variables in a random vector are grouped together because there may be correlations among them; often they represent different properties of an individual statistical unit. The components of the random vector are $c= (c_1, c_2, \ldots c_n)$. The purpose of this process is to convert the given data a into maximally independent components c measured by some function of independence. This function may be like $F(c_1, \ldots c_n)$. It will be done using a linear static transformation as:

$$s = Wx$$

Independent component analysis can be divided into noiseless and noisy cases, where noiseless ICA is a special case of noisy ICA. In the next section noiseless ICA is described.

1. Noiseless ICA

Noiseless independent component analysis can be calculated as follows:

The components a_i of the observed random vector a = $(a_1, a_2, \ldots a_p)^T$ are generated as a sum of the independent components c_k, k=1, 2, ... n:

$a_i = m_{i,1}c_1 + m_{i,2}c_2 + \ldots + m_{i,k}c_k + \ldots + m_{i,n}c_n$

where $m_{i,k}$ are the weights.

The same generative model can be written in vectorial form as , $a = \sum_{k=1}^{n} c_k m_k$ where the observed random vector a is represented by the basis vectors $m_k = m_{1,k}, \ldots, m_{p,k}$. The basis vectors m_k form the columns of the mixing matrix $M = m_1, \ldots, m_n$ and the generative formula can be written as$x = Mc$, where $c = c_1, \ldots, c_n$.

Given the model and samples $a_1, a_2, \ldots a_p$ of the random vector a, the task is to estimate both the mixing matrix M and the sources c. This is done by adaptively calculating the w vectors and setting up a cost function which either maximizes the non-Gaussianity of the calculated $c_k = w^T * a$ or minimizes the mutual information. In some cases, a priori knowledge of the probability distributions of the sources can be used in the cost function.

The original sources c can be recovered by multiplying the observed signals a with the inverse of the mixing matrix $W = M^{-1}$, also known as the unmixing matrix. Here it is assumed that the mixing matrix is square ($n = p$). If the number of basis vectors is greater than the dimensionality of the observed vectors,$n > p$, the task is over-completed.

2. Objective Functions

There are different types of objective functions found in ICA computation. Some of these are as follows:

i. Multi-Unit Objective Functions

There are many different ways to specify objective functions. This section lists several possibilities. It has been shown, that despite their different formulations, they all closely related, and under certain conditions, some are equivalent. Cumulant and general contrast-based methods, however, can be used for any non-Gaussian data, without knowing the underlying distributions. Objective functions of this category are:

- *Maximum likelihood and network entropy*
- *Mutual information and Kullback-Leibler divergence*
- *Non-linear cross-correlations*
- *Non-Linear PCA*
- *Higher-order cumulant tensors*

ii. One-Unit Objective Functions

One-unit contrast functions [8] seek a single vector *w* such that the linear combination $x^T w$ is equal to one of the independent components s_i. It is desirable when not all the PCs are needed, it can be used iteratively to find more PCs, and it tends to result in computationally simple solutions. Some of the functions of this category are:

- *Negentropy*
- *Higher-order cumulants*
- *General contrast functions*

Independent component analysis was originally developed to deal with problems that are closely related to the cocktail-party problem. Since the recent increase of interest in ICA, it has become clear that this principle has a lot of other interesting applications as well. Another, very different application of ICA is on feature extraction. A fundamental problem in digital signal processing is to find suitable representations for image, audio or other kind of data for tasks like compression and denoising. Data representations are often based on (discrete) linear transformations. Standard linear transformations widely used in image processing are the Fourier, Haar, cosine transforms etc. Each of them has its own favourable properties (Gonzales and Wintz, 1987).

C. Discrete wavelet transform

Discrete wavelet transform (DWT) is a wavelet transform technique. For this technique the wavelets are discretely sampled. The wavelet transform has become a useful computational tool for a variety of signal and image processing applications. It is a useful technique for the compression of digital image files because smaller image files are important for storing images using less memory and for transmitting images faster and more reliably.

Wavelet transform is basically of two types. One type of wavelet transform is designed to be easily invertible. It means that the original signal can be easily recovered after it has been transformed. This kind of wavelet transform is used for image compression and cleaning. Typically, the wavelet transform of the image is first computed, the wavelet representation is then modified appropriately, and then the wavelet transform is inverted to obtain a new image. The second type of wavelet transform is designed for signal analysis; for example, to detect faults in machinery from sensor measurements.

There are various wavelet transforms available in the literature. Amongst these Haar wavelet transform is the most basic one. It was described by Alfred Haar in 1910. It serves as the prototypical wavelet transform. The basis of the Haar transform is the decomposition of a signal.

When DWT is applied to a data vector D , its transforms the vector to a numerically different vector, D', of wavelet coefficients. The two vectors are of the same length. When this technique is used in data

reduction, each object will be considered as an p-dimensional data vector, that is, $D = (d_1, d_2, \ldots, d_n)$, depicting n measurements made on the tuple from p database attributes. This technique is a great help in data reduction. The wavelet transformed data can be truncated. A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients.

For example [9], all wavelet coefficients larger than some user-specified threshold can be retained. All other coefficients are set to 0 . The resulting data representation is therefore very sparse, so that operations that can take advantage of data sparsity are computationally very fast if performed in wavelet space. The technique also works to remove noise without smoothing out the main features of the data, making it effective for data cleaning as well. Given a set of coefficients, an approximation of the original data can be constructed by applying the inverse of the DWT used.

A general procedure for applying a discrete wavelet transform as described in [9] is as follows:

- The length, L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary ($L \geq n$).
- Each transform involves applying two functions. The first applies some data smoothing, such as a sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.
- The two functions are applied to pairs of data points in X, that is, to all pairs of measurements (x2i , x2i+1). This results in two sets of data of length L/2. In general, these represent a smoothed or low-frequency version of the input data and the high frequency content of it, respectively.
- The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of length 2.
- Selected values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

D. Factor Analysis

Like PCA, factor analysis (FA) [8] is also a linear method, based on the second-order data summaries. FA assumes that the measured variables depend on some unknown, and often un-measurable, common factors. Typical examples include variables defined as various test scores of individuals; as such scores are thought to be related to a common "intelligence" factor. The goal of FA is to uncover such relations, and thus can be used to reduce the dimension of datasets following the factor model. Various factor analysis techniques are:

1) Principal Factor Analysis

As its name suggests, principal factor analysis (PFA) is related to principal component analysis. It seeks the least number of factors which can account for the common variance (correlation) of a set of variables.

2) Maximum Likelihood Factor Analysis

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

III. CONCLUSION

From the study it is found that analysis on the strength and weakness of PCA is that it is a non-parametric analysis. There are no parameters to twist and no coefficients to adjust based on user experience - the answer is unique and independent of the user. This same strength can also be viewed as a weakness. If one knows a-priori some features of the dynamics of a system, then it makes sense to incorporate these assumptions into a parametric algorithm or an algorithm with selected parameters.

From the study it is found that ICA is a general-purpose statistical technique in which observed random data are linearly transformed into components that are maximally independent from each other, and simultaneously have interesting" distributions. ICA can be formulated as the estimation of a latent variable model. The intuitive notion of maximum nongaussianity can be used to derive different objective functions whose optimization enables the estimation of the ICA model.

The discrete wavelet transform provides sufficient information both for analysis and synthesis of the original signal, with a significant reduction in the computation time. The DWT is considerably easier to implement.

Again factor analysis is a method for investigating whether a number of variables of interest are linearly related to a smaller number of unobservable factors or not.

REFERENCES

[1] T. Jolliffe , Principal Component Analysis, Springer Series in Statistics, Edition 2, Springer, 2002, ISBN 0387954422, 9780387954424.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed. Academic Press, 1990.

[3] Hotelling, H., 1936: Relations between two sets of variants. Biometrika, 28, 321-377.

[4] Aapo Hyvärinen and Erkki Oja, Independent Component Analysis: Algorithms and Applications, Neural Networks

Research Centre, Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland, Neural Networks, 13(4-5):411-430, 2000

[5] Article available online at http://en.wikipedia.org/wiki/Multivariate_random_variable

[6] Article available online at http://en.wikipedia.org/wiki/Principal_component_analysis

[7] Article available online at http://en.wikipedia.org/wiki/Independent_component_analysis

[8] Imola K. Fodor, A survey of dimension reduction techniques, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory Livermore, CA, Material Available Online at: https://computation.llnl.gov/casc/sapphire/pubs/148494.pdf

[9] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Second Edition, 2009, ISBN: 978-81-312-0535-8