Survey on Feature Selection and Dimensionality Reduction Techniques

Govinda.K[1], Kevin Thomas [2]

[12]*School of Computing Science Engineering, VIT University, Vellore, India*

---***---

Abstract - *Data mining methods are used to extract useful information from huge amount of data. Mining of data is a time consuming process because the data that has to be mined may be comprised of large number of dimensions. So we can say that number of dimensions exponentially vary with computation time. In order to speed up the decision making process, dimensionality reduction techniques came into existence. In this paper we have discussed about various techniques of dimensionality reduction that exists currently.*

Key Words: **PCA, SVD, SVM, LCM, CCA**

1. INTRODUCTION

Advancement in data collection and data storage during the past decade has led to an information overload. Researchers from the field of engineering, biology, astronomy, remote sensing, consumer transaction and economics face larger observations and simulation every day. [1] Data mining aims at processing, classifying and selecting useful data from any given dataset. The dataset on which data mining is done may consist of information of varied form and enormous size. This information may be inappropriate or may show missing values that may mislead the user due to the irregularities. These irregularities are detected, analyzed and modified during data pre- processing by applying certain modifications and behavioral predictions.

Dimensionality reduction is among the predominant techniques of data pre-processing which helps to incorporate a systematized structure into the dataset, prior to the mining process. It is also an effective way to downsize data which encourages effective storage and retrieval of data. Some of its important applications are found in the areas like text mining, image processing, pattern recognition, image retrieval etc. Dimensionality reduction is implemented using various algorithms such as SVD, PCA, ICA, SVM etc. [2] This paper describes and analysis OF nine such algorithms to find each ones strength and weaknesses.

1.1 Why Dimensionality Reduction

- It is easy and convenient way to collect data
- Data that is collected, is not just from data mining
- Data is accumulated in an unprecedented speed
- For effective machine learning and data mining data processing is an important part
- Helps in downsizing data
- Visualization
- Data compression
- Noise removal

1.2 Applications of Dimensionality Reduction

- Text mining
- Image retrieval
- Customer management
- Face recognition
- Intrusion detection
- Microarray data analysis
- Handwritten digit recognition
- Protein classification

2. MAJOR TECHNIQUES OF DIMENSIONALITY REDUCTION

2.1 Feature Selection

A process that chooses an optimal subset of features according to an objective function [3]. Feature selection is done to remove noise, reduce dimensionality and to improve mining performance such as speed of learning, predictive accuracy and to make the mined results simple to understand.

2.2 Feature Extraction

It refers to mapping of high-dimensional data to a lower-dimensional space. [3]. Criterion for this may differ based on different problem settings such as:

- Unsupervised Setting – minimum loss of information
- Supervised Setting – maximum class discrimination

3. ALGORITHMS

3.1 Singular Value Decomposition (SVD)

SVD is a gene selection procedure that is performed to reduce dimensionality in data. It is a matrix factorization method which comes under linear vector algebra. The main objective of SVD in data analysis of gene expression is to identify and extract the structural constraints within the data and also to relate significant associations. The main idea behind SVD is to calculate eigenvalues and eigenvectors of covariance matrix of sample-gene matrix [4]. The eigenvalues would help to deduce the respective (or corresponding) eigenvectors; for higher eigenvalues the corresponding eigenvectors would contain a higher variability. Usually the eigenvectors that appear initially with higher unpredictability are considered as prime Principle Components (PCs). These PCs are used to reduce the data into smaller dimensions [5].

The tool used to analyze the data is known as clustering. Group of genes the show similarity (i.e. similar expression profile) are discovered by Clustering. The two methods in clustering are Model-Free andModel-based methods[6]. Model-freemethods do not have any probabilistic tree whereas model-based clustering method clusters are build based on assumptions such that data follows mixture distribution.

SVD overcomes two main difficulties in gene expression data that is parameter estimation and normality assumptions on gene expression. After applying SVD on data a technique called probit transformation is done to improve the working of model-freemethod. SVD can be applied to dataset that has both scattered and un-scattered genes.

3.2 Partial Least Squares Regression (PLSR)

Research in science and engineering often involves using controllable and/or easy-to-measure variables (factors) to explain, regulate, or predict the behavior (responses) of other variables. When the factors are few in number, they are not significantly redundant (collinear), and have a well-understood relationship to the responses, then a good way to turn data into information could be multiple linear regression (MLR). When any one of these three conditions does not satisfy, MLR proves to be inefficient or inappropriate. Hence, researchers face problems with many variables and ill-understood relationships, while constructing a good predictive model. Example: Spectrographs.

When the factors are many and highly collinear we use Partial Least Squares (PLS) method for constructing predictive models. This emphasizes on predicting the response rather than trying to understand the relationship between the variables. For example, PLS is usually not appropriate for screening out factors that have a negligible effect on response. However, when the goal is prediction and there is no practical need to limit the number of factors, PLS may be a very useful tool.

MLR can be used with many factors. However, if the number of factors gets too large, you are likely to get a model that fits the sampled data perfectly but that will fail to predict new data well. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there may be only a few underlying factors or latent factors that account for most of the variation in response. The main intention of PLS is to extract these latent factors, accounting for as much of the manifest factor variation as possible while modelling the responses well[7].

3.3 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) is a method which finds its application in statistics, machine learning and pattern recognition to find a linear combination of features which characterizes or separates two or more classes of objects. This resulting combination is used as a linear classifier and also for dimensionality reduction before later classification.

LDA is closely related to ANOVA (analysis of variance) and regression analysis, attempts to express one dependent variable as a linear combination of other features. However, ANOVA uses categorical independent variables and a continuous dependent variable. In discriminant analysis continuous independent variables and a categorical dependent variable i.e. the class label are used. Logistic regression is more similar to LDA, as it explains a categorical variable by the values of continuous independent variables. We prefer these methods in applications where it is not reasonable to assume that the independent variables are normally distributed. This is the fundamental assumption of the LDA method.

When the measurements made on independent variables for each observation are continuous quantities LDA is used. For categorical independent variables, a similar technique is discriminant correspondence analysis [8].

3.4 Locally Linear Embedding (LLE)

High-dimensional data can often be converted to low-dimensional data with little or no fundamental loss of information. A simple and widely used method for dimensionality reduction is principal component analysis (PCA). In this method data points by their respective orthogonal projections on a subspace of low dimension spanned by the directions (also called components, features, factors or sources) of greatest variance in the data set is represented. The two recently proposednon-linear generalizations of PCA are locally linear embedding (LLE) and ISOMAP algorithms. It was developed for visualization purposes, these two methods projecthigh-dimensional data into a two or low-dimensional subspace by extracting meaningful components in a non-linear fashion. The locally linear embedding algorithm assumes that a high-dimensional data set lies on, or near to, a smooth low-dimensional manifold. Small patches of the manifold, each of which contains a fraction of the data set, can be equipped with individual localco-ordinates. High-dimensional co-ordinates of each patch can be mapped into corresponding local co-ordinates by means of an essentially linear transformation. This method attempts to find a global transformation of the high-dimensional co-ordinates into low- dimensional ones by exploiting adjacency information about closely located data points, this information being a form of summarization of the local transformations between the high- and low-dimensional co-ordinates[9].

3.5 Generic Algorithm

Computational studies of Darwinian evolution and natural selection have led to numerous models for resolving optimization. Generic Algorithms comprise a subset of these evolution based optimization problems techniques focusing on the application of selection, mutation, and recombination to a population of competing problem solutions. Generic Algorithms are parallel, iterative optimizers, and have been successfully applied to a broad spectrum of optimization problems, including many pattern recognition and classification tasks [10].

3.6 Principal Component Analysis (PCA)

It is a Feature Extraction technique that is used to analyze 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 [11]. 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.

Simple Principal Components Analysis technique take a data oriented approach and it is used in case of dimensionality reduction of two highly dimensional picture databases. In such cases it has showed quick convergence rate and stability and has produced estimated solutions without considering thevariance-covariance matrix [12]. So this technique is more beneficial in use as compared to other existing methods.

Matrix method and data method are the key methods for PCA. In the matrix method as the name suggest, matrix is formed for the datasets which comprises of the calculated variance and co variance of the data. Further diagonalization technique is also applied on the matrix. Data methods directly deal with the data. PCA is generally applied before clustering of the datasets happens so as to get an improved drawing out of the cluster organization.

This technique also has a disadvantage of not considering the class separability because of the absence of class label of the feature vector [13].PCA mainly focuses on the rotation of the coordinate axes, that is transformed in the direction which as the maximum variance.

The traditional PCA is implied only on linear transformation and is not effective fornon-linear data. So, to overcome this drawback, non-linear transformation is applied to high dimensional space [14].

3.7 Support Vector Machines (SVM)

SVM is a technique which can be applied to several domains consecutively. SVM mainly focus on the microarray cancer data that contains huge number of gene expressions. The objective of SVM is to get the most optimal separating hyper-plane that has the maximum margin (w).this techniques is more suitable in analyzing gene expressions because it specifically concentrate on datasets which has lower number of samples as compared to the genes and it is also useful in gene selection.

SVM is useful because the results produced by this technique are optimal and distinctive. It is s used in combination with Recursive Feature Replacement (RFR) that produces a unique assessment of generalization of feature (genes).it starts with the gene set which is complete; it removes the least significant gene from the classifier in subsequent iterations. There is another method called SVM-T-RFE method which works on smaller gene subsets and thus produces the highest accuracy.

There is one more method similar to SVM, it is called as SVD. The major difference between the two methods is that SVM is a supervised method while SVD is an unsupervised method. They work on the same application areas like feature selection and modelling of dynamics of gene expression and clustering. But SVM is considered to be more effective over other existing procedures [15].

3.8 Independent Component Analysis (ICA)

The computational technique that is used to obtain reduced subcomponents from the assorted signal after splitting them is called Independent Component Analysis technique.to understand this method we can consider a simple example of ICA called “cocktail party ”. Here a sample data is taken which consists of individuals interacting with each other in a room and there fundamental speech signals are divided. In a generalized way we can say for N sources at least N estimations are needed for primal signals mining.

This can be also achieved using Regularized Whitening method in which dimension of independent sets is reduced to limited set. This algorithm has the potential to incorporate statistics of higher order that comprises of complementary data.it finds its importance when sending of data differs from its prevalent parameters. Whitened data is needed for this method, which is obtained from the identity covariance matrix. It can be used as an augmented version of the PCA method. Initial projection vectors are used in ICA, using thee parameters, three more algorithms are developed for ICA-DR. This procedure is further classified into three subprocedures. ICA-DR1, is the first procedure which uses virtual parameter together with prioritization and selection components. The next one is the ICA-DR2, here ICA is executed as an unauthorized algorithm and its randomness is characterized by randomly choosing initial projection vectors. The third algorithm is called the ICA-DR3, this method uses initialization algorithm together with virtual dimension to develop more appropriate initial projection vectors group.so that they can be used to substitute these vectors required by ICA to develop each of its ICs.

ICA is closely related to Projection Pursuit (PP), which is another statically data analysis tool, whose aim is to reduce the dimensions of the multivariate data sets by identifying the"interested projections "from the projection directory.

3.9 Canonical Correlation Analysis (CCA)

This technique was first introduced by Harold Hoteling; he used cross covariance matrices for dimensionality reduction.in this technique, two different set of variables are taken along with their correlations, this correlations helps to establish linear combinations of the sets with the set of highest correlation. CCA is used in the creation of a model equation that is used to relate two set of variables, for example one set comprises of performance measures and the other one consists of descriptive variables.

CCA is also used for class prediction taken from standard classes that has primal set of samples. This method along with its regularized version(RCCA) are used for blending two modalities.it is mainly used for establishing linear relationship between the text associated to an image and the pixel values of the same image. RCCA is used to study gene expressions in liver cells and finally it links this data with the concentrated hepatic acid in mice. CCA has disadvantage of not fitting into modalities which have large quantities of dimensions.RCCA overcome this drawback but is a bit expensive method. It also has a disadvantage of regularization and thus produces defectives inverses. This is also overcome by RCCA but failed to create equivalent level of refinement between patient classes.

CCA is used to design linear affiliation that comprises of multidimensional variables.it chooses two bases ideal with correlations and these bases are assigned to each variable. Then it selects two bases from the correlation matrix in which diagonal elements are variables and whose correlation is maximum. This is the only difference between ordinary correlation analysis and the CCA.

4. CONCLUSIONS

The objective behind the survey is to provide a complete understanding of the various algorithms used in dimensionality reduction and to analyze the developing interest in this field during the past few years. Dimensionality reduction technique based on dictionaries and projections are growing rapidly. PCA is the most preferred technique due to its simplicity. In middle 2000's ICA was at its boom but later it started seeing its decline. But it will continue to in demand for the applications related to single processing. The rest of the techniques have placed themselves well in the market. Overall we may conclude that dimensionality reduction techniques have been and will continue to be applied in many sectors ranging from biomedical research to pattern recognition. In this survey paper we have covered different methodologies; each requiring different criteria but all having the same goal of reducing the complexity at the same time to deliver a more appropriate (understandable) form of the information.

REFERENCES

[1] A survey of dimension reduction techniques Imola K. Fodor Center for Applied Scienti¯c Computing, Lawrence Livermore National Laboratory P.O. Box 808, L-560, Livermore, CA 94551.

[2] A SURVEY OF DIMENSIONALITY REDUCTION AND CLASSIFICATION METHODS, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3, June 2012

[3] Lei Yu, , Jieping Ye, Huan Liu , "Dimensionality Reduction for Data Mining Techniques, Applications and Trends; Binghamton University and Arizona State University.

[4] Abdi, H. Lewis-Beck, M.; Bryman, A. & Futing, T. , Encyclopedia for research methods for the social sciences Factor rotations in factor analyses Sage, 2003, 792-795.

[5] Abdi, H. Salkind, N. (ed.) Encyclopedia of measurements and statistics Singular value decomposition (SVD) and Generalized Singular Value Decomposition (GSVD) Sage Publications, 2007, 907-912.

[6] Aharon, M.; Elad, M. & Bruckstein, A. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation IEEE Trans. Signal Processing, 2006, 54, 4311-4322.

[7] CSCE 666 Pattern Analysis | RicardoGutierrez-Osuna | CSE@TAMU .

[8] Modern Methods For Business Research edited by George A.Marcoulides . An Introduction to Genetic Algorithms Mitchell Melanie A Bradford Book The MIT Press

[9] Cambridge, Massachusetts • London, England Fifth printing, 1999 First MIT Press paperback edition, 1998 Copyright © 1996 Massachusetts Institute of Technology.

[10] S. Bicciato, A. Luchini, C. Di Bello, "Disjoint PCA Models for Marker Identification And Classification Of Cancer Types Using Gene Expression Data,"2002, IEEE 0-7803-7557-2.

[11] Matthew Partridge, Rafael Calvo, "Fast Dimensionality Reduction and Simple PCA,"2006.

[12] Vinay Nadimpally , Mohammed J. Zaki,"A Novel Approach to Determine Normal Variation in Gene Expression Data," SIGKDD Explorations, Vol. 5 Issue 2,2002.

[13] Isabelle Guyon, Jason Weston, Stephen Barnhill, M.D., "Gene Selection for Cancer Classification using Support Vector Machines," Barnhill Bioinformatics 2001.

[14] Jing Wang, Student Member, IEEE, andChein-I Chang, Senior Member, "Independent Component Analysis Based Dimensionality Reduction with Applications in Hyper spectral Image Analysis," IEEE Transactions On Geoscience And Remote Sensing Vol. 44, 2006, IEEE 0196-2892.

[15] Abhishek Golugula, George Lee, Stephen R. Master, Michael D. Feldman, John E. Tomaszewski, and Anant Madabhushi, "Supervised Regularized Canonical Correlation Analysis: Integrating Histologic and Proteomic Data for Predicting Biochemical Failures," EMBS 2011, IEEE 978-1-4244-4122-8.

BIOGRAPHIES

Dr.K.Govinda received the degree in computer science and engineering from Nagarjuna University in 1998. He is Associate Professor in School of Computing Science and Engineering, VIT University. His interests are database, data warehousing and cloud computing.