A Survey of Data Mining Techniques for Dimensionality Reduction of High-Dimensional Data

Rahul Patidar
M Tech Scholar
VITS Bhopal, India
E-mail- patidar_rahul@rediffmail.com

Sumit Sharma
Assistant Professor
VITS Bhopal, India
E-mail- sharma782022@yahoo.co.in

Abstract
Dimensionality Reduction (DR) methods are a cornerstone of analyzing high dimensional data, due to their simple geometric interpretations and typically attractive computational properties. These methods capture many data features of interest, such as covariance, dynamical structure, and correlation between data sets, input-output relationships, and margin between data classes. In this paper we survey about methods from this disparate literature as optimization programs over matrix manifolds. We discuss about Singular Value Decomposition (SVD), principal component analysis (PCA), Support vector machines (SVM), Locally Linear Embedding (LLE), Linear discriminate analysis (LDA), canonical correlations analysis (CCA), independent component analysis (ICA), and Partial Least Squares Regression (PLS REGRESSION). Our motivation is not to be comprehensive but to present summary of basic techniques, as well as to review select state-of-the-art methods. In this survey paper we give introduction to dimension reduction from a visualization point of view.

Keyword: - ***Dimensionality reduction, high-dimensional, manifold learning, feature extraction, Cluster analysis, High-dimensional categorical data.***

INTRODUCTION

Contemporary simulation and experimental data acquisition technologies enable scientists and engineers to generate massive amounts of data. Thereby, more and more application domains are producing progressively larger and inherently more complex (multivariate) data sets. These data sets are collections of samples that consist of multiple measured (or simulated) observations of a variable set. Expressed in a space that requires many degrees of freedom multivariate data present severe problems for data analysis and especially for visualization. Visualization is the integral part of exploratory data analysis, the first stage of data analysis where the goal is to make sense of the data before proceeding with more goal-directed modeling and analyses. Since human perception (and output devices) is limited to three-dimensional space, the challenge of visualizing multivariate data is converting the data to a space of lower dimensionality that is depictable and comprehensible to the user while preserving as much information as possible. This process is called dimension reduction and visualization of multivariate data is one of its traditional applications.

This survey reviews methods of dimension reduction that focus on visualizing multivariate data. That is, they are suitable for a depictable target space. Thereby, we describe the concepts and ideas underlying the algorithms. Implementation details, although important, are not discussed. The reader should be aware that there are numerous dimension reduction methods that focus on the various aspects of data analysis. For example, methods for feature reconstruction or classifications are closely related to those considered here, but are not discussed because their focus is not visualization. The reader will find that, due to its long history, there are numerous surveys on dimension reduction. For example, authors focus on a specific subset of techniques or investigations; provide a broad overview, or historical background. This survey provides an introduction to the concepts of visualizing high-dimensional data using dimension reduction and reviews select state-of-the-art methods that share this focus.

The remainder of the paper is structured as follows. Section II represents the core of the clustering. Section III reviews valuable work done previously by researchers. In Section IV we give some ideas for possible future research and finally we conclude our paper in Section V.

II CLUSTERING

Data clustering (or just clustering), is an unsupervised classification technique, whose ultimate aim is to creating groups of objects, or clusters, in this way that objects in the same cluster are very similar and objects in different clusters are quite distinct [1]. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover interests of their customers based on purchasing patterns and characterize groups of the customers. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In geology, specialist can employ clustering to identify areas of similar lands, similar houses in a city and etc. data clustering can also be helpful in classifying documents on the Web for information discovery [2].

Data clustering, also called cluster analysis, is a challenging field of research in which applications pose their own special requirements. Data mining applications place the following special requirements on clustering techniques [2]:

- **Scalability:** Clustering applications may have a large database that contains millions of objects. So, highly scalable clustering algorithms are needed to successfully form the clusters.
- **Ability to deal with different types of attributes:** data points may have different types such as numerical, ordinal, categorical, and binary. Different applications may require clustering data of a one type of mixture of data types.
- **Arbitrary shape discovery:** Some clustering algorithms determine clusters based on distance measurement such as Euclidean. These algorithms form spherical clusters. Other clustering algorithms are needed to find clusters of arbitrary shapes such as those based on density.
- **Insensitivity to noise**: Clustering algorithms are needed to be insensitive to noise and outlier data to avoid the result of poor clustering.
- **High dimensionality:** Many clustering algorithms can efficiently find clusters of low dimensional data. However, clustering in high dimensional space is a challenging task since distances between objects become very large and average density of points is likely to be quite low.

In the literature, many clustering algorithms have been proposed. These algorithms differ from each other by the criteria considered which lead to different categories of clustering algorithms. Although it is difficult to find strict categorization of the clustering algorithms because the categories may overlap, the following categorization is helpful to discriminate the clustering algorithms [2]:

- **Partitioning methods**: A partitioning method creates k partitions (or clusters) such that $k \leq n$ where n is the total number of objects. It creates an initial portioning and then iteratively moves the objects from one cluster to another to improve the partitioning. Good clustering is that the similarity between objects in the same cluster is high whereas the dissimilarity between objects in the different clusters is high. The k-means algorithm is a commonly used partitioning method [3].

- **Hierarchical methods:** A hierarchical method creates a hierarchal structure of the data objects. Then a given number k of clusters determines how to cut the hierarchy. It can be either agglomerative or divisive. The agglomerative approach starts by considering each object as a separate cluster. Then it iteratively merges the most similar clusters until grouping all the objects in one cluster. The divisive approach starts by considering the entire objects as one cluster. Then it iteratively splits each cluster into smaller clusters until each objects forms its own cluster. AGNES and DIANA [4] are examples of hierarchical clustering. BIRCH [5] integrates hierarchical clustering with iterative (distance-based) relocation.
- **Density-based methods:** The idea behind these methods is to group dense objects into clusters. An object is dense if its neighborhood if a given clusters contains at least minimum number of objects. These methods have the advantage to find clusters with arbitrary shapes and they are insensitive to noise and outliers. DBSCAN [6] and OPTICS [7] are typical examples of density-based clustering.

- **Grid-based methods:** These methods divide the object space into a finite number of cells that form a grid structure. Therewith connected cells are grouped in a cluster. STING [9] is an example of grid-based clustering. Some techniques such as CLIQUE [10] combine both density-based and grid-based approaches. The main advantages of these methods are fast processing and arbitrary-shape clusters foundation.

- **Model-based methods:** This approach creates a mathematical model for each of the clusters and finds the best fit of the data to the given model. A main advantage is that these methods automatically determine the number of clusters based on standard statistics. COBWEB [11] and self-organizing feature maps [12] are examples of model-based clustering.

- **Methods for high-dimensional data:** Distance- and Density-based methods are inefficient for clustering high-dimensional data since objects are increasingly sparse. Alternative approaches, such as subspace clustering methods and frequent pattern-based clustering, have been proposed. Subspace clustering methods search for clusters in subspaces of the data, rather than over the entire data space. CLIQUE [10] and PROCLUS [13] are examples of subspace clustering methods. Frequent pattern-based clustering methods extract distinct frequent patterns among subsets of dimensions that occur frequently. P Cluster [8] is an example of frequent pattern-based clustering that groups objects based on their pattern similarity.

BASIC CONCEPTS OF CLUSTER ANALYSIS

In this section, we introduce some basic concepts that are frequently encountered in the field of cluster analysis, i.e. objects and attributes, similarity and dissimilarity, dataset, and cluster centers.

Dataset

A dataset is a collection of data items that have different characteristics. In clustering, these data items are grouped into clusters.

Objects and Attributes
An object is a single data item, i.e. a member in a dataset. It can also be referred to as data point, pattern case, observation, individual, item, tuple, record, or object. An attribute is value that specifies of a property of the object such as length, weight, etc. It can also be referred to as variable, or feature.

Data Types
Data clustering algorithms are associated with types of data the attributes have [1]. An attribute can be Binary, Categorical, Ordinal, Interval-scaled, or Ratio-Scaled [2].

- **Binary attributes** has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present.
- **Categorical attributes,** also referred to as nominal, are simply used as names, such as the brands of cars and names of bank branches. That is, a categorical attribute is a generalization of the binary variable; it can take on more than two states.
- **Ordinal attributes** resembles a categorical variable, except that the M states of the ordinal value are ordered in a meaningful sequence. For example, professional ranks are often enumerated in a sequential order, such as assistant, associate, and full for professors.
- **Interval-scaled attributes** are continuous measurements of a linear scale such as weight, height and weather temperature.
- **Ratio-Scaled attributes** make a positive measurement on a nonlinear scale. For example an exponential scale, , and the volume of sales over time are ration-scaled attributes.

Similarity and Dissimilarity
In data clustering, similarity and dissimilarity measures are used to describe quantitatively the similarity or dissimilarity of two data points or two clusters. Similarity coefficients are used to describe quantitatively how similar two data points are or how similar two clusters are: the greater the similarity coefficient, the more similar are the two data points. Dissimilarity measure, such as distance, is the other way around: the greater the dissimilarity measure or distance, the more dissimilar are the two data points or the two clusters. For example, the Euclidean distance between two objects **x** and **y** is considered a dissimilarity measure.

Cluster and Cluster Center
In data clustering, a cluster is a group of objects that have common properties, show small dissimilarities, have relations with at least one object in the cluster, and are clearly distinguishable from the rest of objects in the dataset. A cluster center is a reference point in the cluster. That is, the center is the representative of the cluster. Figure 1 shows three well-separated clusters, each of them is represented by its center.

Hard Clustering and Fuzzy Clustering
In hard clustering, algorithms assign a class label li {1, 2. . . k} to each object xi to identify its cluster class, where k is the number of clusters. In other words, in hard clustering, each object is assumed to belong to one and only one cluster. Mathematically, the result of hard clustering algorithms can be represented by a k×n matrix. In fuzzy clustering, the assumption is relaxed so that an object can belong to one or more clusters with probabilities.

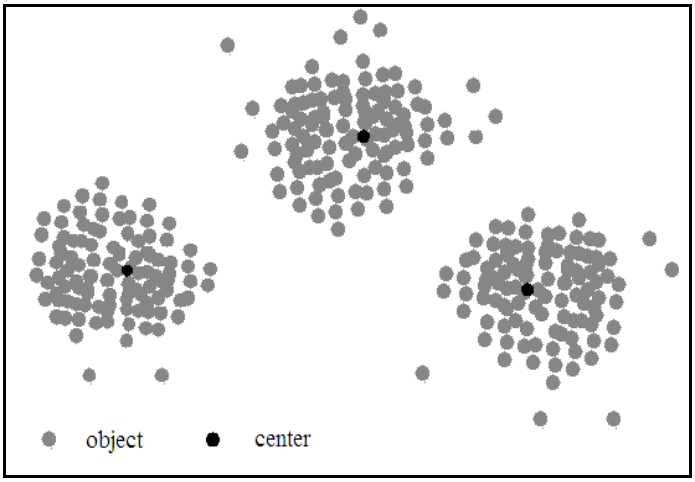

Figure 1: Three well-separated clusters.

III RELATED WORK
Functions of data mining are association, correlation, prediction, clustering, classification, analysis, trends, outliers and deviation analysis, and similarity and dissimilarity analysis. Clustering technique is applied when there is no class to predict but rather when the instances divide into natural groups [14]. Clustering for multidimensional data has many challenges. These are noise, complexity of data, and data redundancy. To mitigate these problems dimension reduction needed. In statistics, dimension reduction is the process of reducing the number of random variables. The process classified into feature selection and feature extraction [15], and the taxonomy of dimension reduction problems [16] shown in Figure 2. Dimension reduction is the ability to identify a small number of important inputs (for predicting the target) from a much larger number of available inputs, and is effective in cases when there are more inputs than cases or observations.

Dimension reduction methods associated with regression, additive models, neural network models, and methods of Hessian [17], one of which is the local dimension reduction (LDR), which is looking for relationships in the dataset and reduce the dimensions of each individual, then using a multidimensional index structure. Nonlinear algorithm gives better performance than PCA for sound and image data [18], on the other studies mentioned Principal Component Analysis (PCA) is based on dimension reduction and texture classification scheme can be applied to manifold statistical framework.

In most applications, dimension reduction performed as pre-processing step [19], performed with traditional statistical methods that will parse an increasing number of observations [17]. Reduction of dimensions will create a more effective domain characterization [20]. Sufficient Dimension Reduction (SDR) is a generalization of nonlinear regression problems, where the extraction of features is as important as the matrix factorization [21], while SSDR (Semi-supervised Dimension Reduction) is used to maintain the original structure of high dimensional data [22].

Goals of dimension reduction methods are to reduce the number of predictor components and to help ensure that these components are independent. The method designed to provide a framework for interpretability of the results, and to find a mapping F that maps the input data from the space <d to lower dimension feature space <d denotes as F(x) : <d ! <d. Dimension reduction techniques, such as principal component analysis (PCA) and partial least squares (PLS) can used to reduce the dimension of the microarray data before certain classifier is used [23].

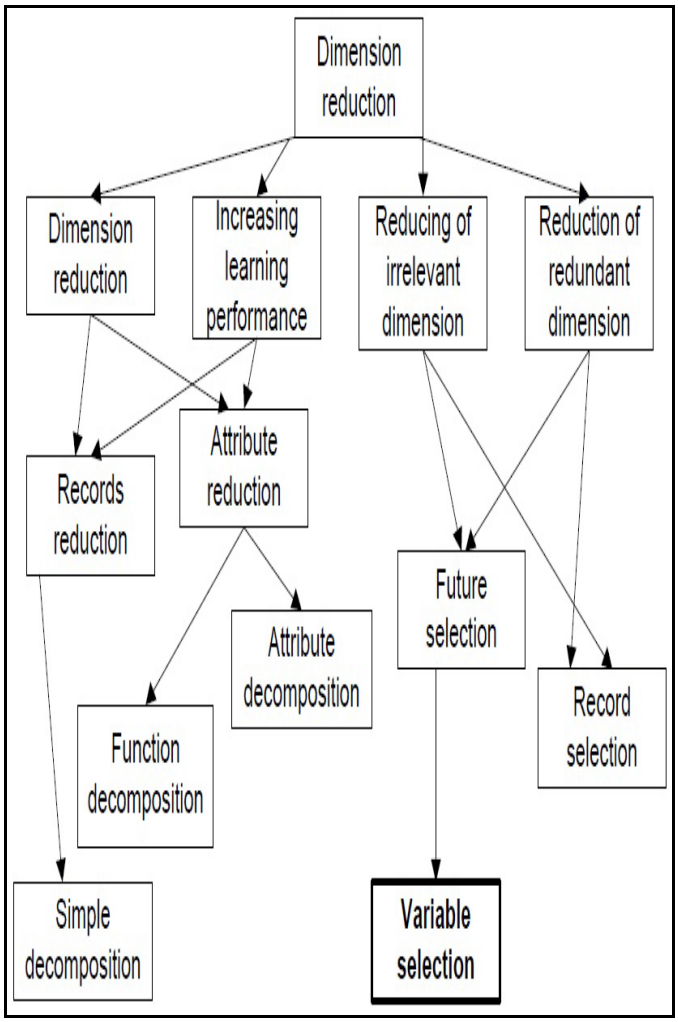

Figure 2: Taxonomy of dimension reduction problem.

Comparison of Dimension Reduction techniques are shown below:

Name of technique	Method	Usage statistics	Working Mechanism
Singular value decomposition (SVD)	Unsupervised	Most of these algorithms has the primary objectives of applying it are identification and extraction of structural constitution within the data.	A gene selection procedure have been used (it is factorization method). [24]
Principle component analysis (PCA).	Unsupervised	Most of these algorithms has Feature extraction objectives	The main focus is to convert the values into set of linear combinations. [25]
Support vector machines (SVM)	Supervised	Most of these algorithms will be specifically appropriate for certain specific datasets where the number of samples is much smaller in comparison to the number of features(genes).	Its main purpose is the analysis of gene expression data. [26]
Independent Component Analysis (ICA)	Unsupervised	Most of the ICA algorithms will require whitened data by means of an identity covariance matrix.	It works on the principle of additive subcomponents and separation of singular units from a large multivariate source. [27]
Canonical Correlation Analysis (CCA)	Unsupervised	Most of CCA algorithms are used DNA micro array assess expressions of numerous thousand of genes	Its main focus is on discovering linear combinations from two sets of variables and then by approximates correlations amongst these variables. [28]
Locally Linear Embedding (LLE)	Unsupervised	These algorithms have been used for its computational plainness and impulsive approach	It works on manifold learning methodology and fall under the non linear practices of dimensionality. [29]
Linear discriminate analysis (LDA)	Unsupervised	Most of these algorithms are used in small sample size problems (for feature selection and feature transformation).	Classification based approach. [30]
Partial Least Squares Regression (PLS REGRESSION)	Unsupervised	PLS discovers a linear regression model by predicting the estimated variables and the perceivable variables to an alternate space.	mathematical approach. [31]

Table 1: Comparison of Dimension Reduction techniques.

IV Research Scope:

There are a number of issues that need further exploration.

- First, fusion of Dimension reduction methods from additional dimensionality reduction methods could be investigated.
- Second, selecting different number of dimensions for different dimensionality reduction methods when fusing classifiers of features could be investigated as an alternative strategy.
- And finally, more sophisticated strategies could be considered to obtain better results.

V Conclusion

The exclusive intention of this survey was to give a vibrant investigation on different trendy and largest algorithmic approaches that are used to execute dimensionality reduction. The Singular Value Decomposition (SVD) method emphasizes on a gene selection technique whereas the Principal Component Analysis (PCA) method focuses on converting the values into a set of linear combinations. The Support Vector Machines (SVM) (supervised) is compared to the Singular Value Decomposition method (unsupervised) and thus their efficiencies are discussed within their applicable expanses.

PCA and Linear Discriminate Analysis (LDA) also fall under a peculiar category of feature transformation where in the former uses a statistical signal criterion whereas the latter uses a classification model. The Partial Least Squares (PLS) method can also be categorized under the same roof of transformation and is compared to PCA where in the former uses a linear regression model whereas the latter stresses on the use of maximum variance calculated. The Locally Linear Embedding (LLE) technique is a manifold learning methodology and thus falls under the non-linear practices of dimensionality reduction.

The Canonical Correlation Analysis (CCA) technique abets in discovering linear combinations from two sets of variables and thereby approximates correlations amongst these variables. The Independent Component Analysis (ICA) method works on the principle of additive subcomponents and separation of singular units from a large multivariate source. Thus dimensionality reduction algorithms can be executed onto a specific dataset with a particular problem, depending upon their usage statistics and parameters under which their applicable conditions are satisfied.

REFERENCES

[1] G. Gan et. al, Data Clustering Theory, Algorithms, and Applications, Siam, 2007.

[2] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd edition, Elsevier, 2006.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observation". In: Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, pp. 281–297.

[4] L. Kaufman and P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

[5] T. Zhang et. al. "BIRCH: An efficient data clustering method for very large databases". In Proceedings of the 1996 ACM SIGMOD international conference on management of data, pp. 103–114. New York: ACM Press, 1996.

[6] M. Ester et. al., "A density-based algorithm for discovering clusters in large spatial databases with noise," In Second international conference on knowledge discovery and data mining", pp. 226–231. Portland, OR: AAAI Press, 1996.

[7] M. Ankerst et. al, "OPTICS: Ordering points to identify the clustering structure," In Proc. 1999 ACM-SIGMOD Int. Conf.Management of Data(SIGMOD'99), pp. 49–60, Philadelphia, 1999.

[8] H.Wang et. al., "Clustering y pattern similarity in large data sets," Proc. ACM SIGMOD Int. Conf. Management of Data, pp. 394–405, 2002.

[9] W. Wang et. al, "STING: A statistical information grid approach to spatial data mining," In Twenty-third international conference on very large data bases, pp. 186–195, 1997.

[10] R. Agrawal et. al, "Automatic subspace clustering of high dimensional data for data mining applications," In SIGMOD Record ACM Special Interest Group on Management of Data, pp. 94–105. New York: ACM Press. 1998.

[11] D. Fisher. "Improving inference through conceptual clustering," In Proc. 1987 Nat. Conf. Artificial Intelligence (AAAI'87), pp. 461–465, Seattle,WA, 1987.

[12] T. Kohonen, "The self-organizing map," Proceedings of the IEEE, 78(9):1464–1480, 1990.

[13] S. Redmond and C. Heneghan, "A method for initializing the k-means clustering algorithm using kd-trees," Pattern Recognition Letters, vol. 28, issue 8, pp. 965–973, 2007.

[14] Sembiring, Rahmat Widia, Jasni Mohamad Zain, Abdullah Embong.Clustering "High Dimensional Data Using Subspace And Projected Clustering Algorithm", International Journal Of Computer Science & Information Technology (IJCSIT) Vol.2, No.4, pp.162-170 (2010).

[15] Nisbet, Robert, John Elder, Gary Miner. Statistical Analysis & Data Mining Application, Elsevier Inc, California, pp.111-269 (2009).

[16] Maimon, Oded, Lior Rokach. Data Mining And Knowledge Discovery Handbook, Springer Science Business Media Inc, pp.94-97 (2005).

[17] Fodor, I.K. A Survey of Dimension Reduction Techniques. LLNL Technical Report, UCRL-ID-148494", pp.1-18 (2002).

[18] Kohonen, Teuvo, Samuel Kaski, and Harri Lappalainen. "Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM." Neural computation 9, no. 6 (1997): 1321-1344.

[19] Ding, Chris, Xiaofeng He, Hongyuan Zha, and Horst D. Simon. "Adaptive dimension reduction for clustering high dimensional data." In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pp. 147-154. IEEE, 2002.

[20] Bi, Jinbo, Kristin Bennett, Mark Embrechts, Curt Breneman, and Minghu Song. "Dimensionality reduction via sparse support vector machines." The Journal of Machine Learning Research 3 (2003): 1229-1243.

[21] Globerson, Amir, and Naftali Tishby. "Sufficient dimensionality reduction." The Journal of Machine Learning Research 3 (2003): 1307-1331.

[22] Zhang, Daoqiang, Hua Zhou Zhi, Songcan Chen' Semi-Supervised Dimensionality Reduction, 7th SIAM International Conference on DataMining, pp.629-634, (2008).

[23] Xu, Rui, Donald C. Wunsch II. Clustering", John Wiley & Sons, Inc, New Jersey, pp. 237-239 (2009).

[24] Faming Liang, "Use of SVD-based probit transformation in clustering gene expression profiles," Computational Statistics & Data Analysis (51) 2007, Science Direct 0167-9473.

[25] Partridge, Matthew, and Rafael A. Calvo. "Fast dimensionality reduction and simple PCA." Intelligent data analysis 2, no. 1 (1998): 203-214.

[26] Krzysztof Simek, Krzysztof Fujarewicz, Andrzej Swierniak, Marek Kimmela, Barbara Garza, Malgorzata Wiench, Joanna Rzeszowska, "Using SVD and SVM methods for selection,classification, clustering and modeling of DNA microarray data," Engineering Applications of Artificial Intelligence (17) 2004, Science Direct 0952-1976.

[27] Liu, Han, and Rafal Kustra. "Dimension Reduction of Microarray Data with Penalized Independent Component Aanalysis." In The NIPS workshop on New Problems and Methods in Computational Biology (NIPS). 2005.

[28] Golugula, Abhishek, 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." In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, pp. 6434-6437. IEEE, 2011.

[29] Y. Goldberg, Y. Ritov, "LLE with low-dimensional neighbourhood representation," Pattern Recognition, 2009.

[30] Edmundo Bonilla Huerta, Beatrice Duval, Jin-KaoHao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," Neurocomputing (73) 2010, ScienceDirect 0925-2312.

[31] Herve Abdi, "Partial Least Squares Regression," 2003.