A Survey of Dimensionality Reduction methods for multidimensional data

Pritika Mehra[1], Mini Singh Ahuja[2]
[1]*Research Scholar, Guru Nanak Dev University Amritsar*
[2]*Assistant Professor, GNDU Regional Campus Gurdaspur*

***Abstract*—Multidimensional data is more prevalent due to the rapid expansion of computational biometric and e-commerce applications. As a result, mining multidimensional data is a crucial issue with significant practical implications. The curse of dimensionality and, more importantly, the meaning of the similarity measure in the high dimension space are two specific difficulties that arise while mining high-dimensional data. The challenges and methods for dimensionality reduction of multidimensional data are surveyed in this work.**

***Index Terms*—Dimensionality Reduction, High Dimensional data, principle component analysis, autoencoders**

I. INTRODUCTION

Nowadays, there is a trend toward more observations, but even more so toward a very large number of variables—the automatic, systematic collection of a significant amount of specific data on each observation. A single observation may have dimensions in the thousands, millions, or billions, but only tens or hundreds of observations are needed for analysis. The observations may be curves, photos, or videos. This data has a high dimensionality. Large datasets and high dimensionality can provide very difficult problems. A wide range of industries, including biometrics, health, e-commerce, network security, and industrial applications, can benefit from high-dimensional data. The right procedures and methods must be used to manage such multi-dimensional data in order to make use of data characteristics. Additionally, data can contain typical properties and multidimensional data structures, which makes it difficult for traditional analysis methods to be effective. Novel ways are needed to analyse more meaningful information from multidimensional data.

II. CHALLENGES IN MULTIDIMENSIONAL DATA

We emphasise the five most difficult problems relating to large data and multidimensional data among the numerous difficult problems:

A. The so-called "curse of dimensionality" is present for high-dimensional datasets, where the searchable volume in hyperspace shrinks relative to the enormous viable search space. Therefore, in order to make sense of the enormous datasets, any solution process can only sample a subset of sparse points with practically zero sampling volume. The task of locating the global optimality is thus extremely difficult and challenging. Additionally, the distance measurements necessary for problem formulation lose some of their significance because any finite distance will provide a ratio of virtually 0 between the massive distance needed to be covered in the high-dimensional search space and the required distance measure.

B. As the number of dimensions rises, the number of features similarly rises, frequently much more quickly, resulting in extremely high levels of sparsity for such high-dimensional features. Furthermore, there can be some link between many dimensions, making it challenging to define characteristics.

C. Datasets for high-dimensional data frequently lack structure, which might make them more difficult to use. Additionally, large datasets frequently contain noise and uncertainties. It can be more difficult to process and apply appropriate data mining tools to such noisy data. There is no analytical strategy for such situations that can offer understanding for even a tiny group of

difficulties. Algorithms are therefore frequently problem- and even data-specific.

D. There are no effective solutions for dealing with such difficult issues since there are an exponentially growing number of possible cluster combinations as the number of dimensions rises (clustering becomes nondeterministic polynomial-time hard, or NP-hard).

E. Despite the fact that current computers are getting faster all the time and there are more affordable parallel and cloud computing options available, high-dimensional information processing still presents a number of difficulties. The need for fresh method and tool development is still very great. In order to solve problems with high-dimensional data, a paradigm shift and unconventional thinking styles may be necessary.

Due to these difficulties, innovative techniques and fresh ideas are required to tackle such complex issues. In reality, it has been found that heuristic and metaheuristic algorithms are a potential class of alternate approaches, particularly those metaheuristic methods based on nature-inspired optimization algorithms.

III. METHODS OF MULTIDIMENSIONAL DATA ANALYSIS

a. Classification

It is a method of supervised learning. It regularly comes up in bioinformatics, such as when classifying diseases using high throughput data from micorarrays or SNPs, and in machine learning, such as when classifying documents and recognising images. From training data made up of pairs of input feature pairs and category output, it attempts to learn a function. Any valid input feature will have its class label predicted using this function. Numerous well-known classification techniques exist, such as Fisher discriminant analysis, support vector machines, (multiple) logistic regression, and k-th-nearest-neighbor classifier. Things get hard when the input feature space has a lot of dimensions.

b. Regression

One of the p variables in a regression setup is a quantitative response variable. Examples include, for instance, the variety of currency rates today given recent exchange rates in financial databases.

I Linear regression modelling: Zi Xi1 Response = a0 + a2Xi2 + • • • + apXip; Xi2,..., Xip predictors; Xi1 = f(Xi2,..., Xip) + Zi is the formula for nonlinear regression modelling (ii).

Latent variable analysis (iii)

It is suggested that in latent variable modelling, X = AS, where X is a vector-valued observable, S is a vector of unobserved latent variables, and A is a linear transformation transforming one into the other The idea is that the basic structure of the array X may be attributed to a small number of underlying latent variables, and that by identifying these variables, we might learn crucial information. One illustration is Principal Component Analysis (PCA). Here, the orthogonal eigenvectors of the covariance matrix C of the observables X are obtained, placed as columns in the orthogonal matrix U, and defined as S = U'X.

Here, A = U is the latent variable form. This method is frequently used in science, engineering, and business applications for data analysis. The projection on the space spanned by the first k eigenvectors of C provides the best rank k mean square approximation to the vector X, which is the mathematical justification for this method.

c. Clustering

Here, the goal is to arrange an unorganised collection of objects in a way that makes objects next to one another comparable. Finding clusters and the space where they are located is the process of clustering high-dimensional data. As a result, there are numerous clustering techniques described in the literature:

Clustering of subspaces

These methods look for clusters in the provided high-dimensional data space's subspaces, where a subspace is defined by a subset of the space's total number of characteristics.

Subspace clustering techniques will look for clusters in a specific data projection [12]. These techniques have the ability to disregard irrelevant features and the correlation clustering problem. Axis-parallel subspaces are regarded as the specific case of two-way clustering, also known as Co-clustering or Biclustering. These techniques cluster the objects simultaneously as the feature matrix made of of data objects spread across rows and [11]. They typically don't function with random feature combinations, as is the case with subspace approaches in general.

The primary approach utilised for numerical attributes for subspace clustering in Quest [13] is CLIQUE-

Clustering. A unit elementary rectangular cell in a subspace serves as the initial point. The cells will be kept if the densities are higher than the specified threshold value [5]. It uses a bottom-up method to locate these components. First, it divides the units into grids of equal-width bin intervals that are one dimension wide. The inputs for this algorithm are threshold and bin intervals [8]. Utilizing the selfjoin of q-1, it proceeds recursively from q-1-dimensional units to q-dimensional units using the Apriori-Reasoning approach. Based on their coverage, the total number of subspaces is ordered. Pruning is done on the less-covered subspaces. A cut point is chosen in accordance with the MDL principle, and a cluster is defined as a collection of connected dense units. It is possible to describe a DNF expression with a finite set of maximal segments, known as regions, whose union equals a cluster [6].

Projected clustering

Although the clusters may be located in several subspaces, projected clustering attempts to associate each point with a certain cluster. The broad strategy combines a standard clustering technique with a unique distance function. According to PROCLUS-Projected Clustering, [2], a tight cluster is associated with a subset of a low-dimensional subspace S such that S is projected into the subspace. A projected cluster will be represented by the pair (subset, subspace). The user will provide inputs for the average subspace dimension n and the number of clusters k [6]. It iteratively locates k medoids, each of which is connected to a certain subspace. The Manhattan distance is utilised to partition the subspace dimension, and a sample of data is also used, coupled with a greedy hill-climbing method.

After the iterative stage is complete, more data passes are made to refine the clusters with subspaces related to the medoids. The earlier proposed PROCLUS technique has been extended to provide ORCLUS-Oriented projected Cluster formation [3]. It makes use of high-dimensional space's projected clustering on non-axes parallel subspaces [9].

Hybrid Clustering

It has been noted on sometimes that not all algorithms attempt to locate a singular cluster for each point, nor are all clusters in all subspaces guaranteed to produce a result in the middle. It's because there are several places that could potentially overlap [7]. Inevitably, the whole sets of clusters are discovered. An elementary method for a subspace clustering algorithm is FIRES [4]. To construct all subspace clusters, it employs a heuristic aggressive strategy [9].

Correlation Clustering

Correlation The feature vector of attribute correlations in a high dimensional space is related to clustering. To direct the clustering process, these are expected to be durable [2]. These correlations may be present in many clusters with various values and cannot be equated to the classical uncorrelated grouping [6]. Different spatial morphologies of clusters are produced by correlations between qualities or subsets of attributes. As a result, the similarity between cluster objects is determined using the local patterns [8]. Given how tightly the two are related, the correlation clustering can be thought of as biclustering. Biclustering will show which sets of items have a correlation in particular qualities. For each cluster, the correlation is typical [10].

The goal of dimensionality reduction techniques is to create a space with a much lower dimension and look for clusters there. A method will frequently combine some dimensions from the original data to create new ones.

IV. DIMENSIONALITY REDUCTION

The technique of translating highly dimensional data to a lower-dimensional embedding is frequently referred to as dimension reduction. Filtering, compression, regression, classification, feature analysis, and visualisation are just a few examples of applications for dimension reduction. Dimensionality reduction is a method for narrowing the feature space in order to create a machine learning model that is stable and statistically sound while escaping the dimensionality curse.

The two primary methods for performing dimensionality reduction are:

Feature transformation and feature selection.

A. The Feature Selection technique aims to delete or subset overlapping or unimportant features while keeping key characteristics.
B. Projecting high-dimensional data into lower dimensions is the goal of feature transformation, also known as feature extraction. Principle

Component Analysis (PCA), Autoencoders, Matrix Factorization, t-Sne, UMAP, and other methods of feature transformation.

Principle Component Analysis

An unsupervised method called Principle Component Analysis projects the original data in the direction of large variance. Because these lines of high variance are orthogonal to one another, the projected data show very little correlation—almost none at all.

Autoencoders

An unsupervised artificial neural network called an autoencoder compresses data to a lower dimension before reconstructing the input again. By concentrating more on the crucial features and eliminating noise and duplication, the autoencoder determines the representation of the data in a smaller dimension. It is built on an encoder-decoder design, in which the encoder converts high-dimensional data to lower-dimensional data and the decoder attempts to convert the lower-dimensional data back to the original high-dimensional data.

V. CONCLUSION

The importance of dimensionality reduction and high dimensionality problems must be urgently addressed. High-dimensional data problems are best solved with high-performance computing techniques. Due to the exponential growth in the dimensionality and sample size, the current algorithms do not always respond in an adequate manner when dealing with extremely high dimension data.

REFERENCES

[1] P. Berkhin, "A Survey of Clustering Data Mining Techniques" Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press, 25-72, 2011.

[2] Guha S., Rastogi R., Shim K,"CURE: An efficient clustering algorithm for large databases", Proc. of ACM SIGMOD Conference, 2012.

[3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2010.

[4] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 2009.

[5] A. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, Volume 31(3), pp. 264-323, 2011.

[6] Zhang T., Ramakrishnan R. and Livny M," BIRCH: An efficient data clustering method for very large databases", In Proc. of SIGMOD96, 2012.

[7] Rui Xu and W. Donald, "Survey of Clustering Algorithms," IEEE Transaction on Neural Network, vol. 16, 2009.

[8] Gan Guojan, Ma Chaoqun, and W. Jianhong," Data Clustering: Theory, Algorithm and Applications", Philadelphia, 2012.

[9] A. Jain and R. Dubes, "Algorithms for Clustering Data", New Jersey, 2011.

[10] A. K. Jain, M. N. Murtyand, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys vol. 31, pp. 264-324, 2012.

[11] K. Bache and M. Lichman. (2013). UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/ml/machinelearningdatabases/

[12] M. Steinbach, L. Ertoz, and V. Kumar, "The Challenges of Clustering High Dimensional Data," in New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, Ed New Vistas: Springer, 2010.

[13] Z. Tian, R. Raghu, and L. Miron, "BIRCH: A New Data Clustering Algorithm and Its Applications," Data Mining and Knowledge Discovery, vol. 1, pp. 141-182, 2009.

[14] Xin-She Yang,Sanghyuk Lee, Sangmin Lee, and Nipon Theera-Umpon, Information Analysis of High-Dimensional Data and Applications, *Mathematical Problems in Engineering(2015)*

[15] [15]*W*ang W., Yang J. (2005) Mining High-Dimensional Data. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_37

[16] https://www.isical.ac.in/~acmsc/WBDA2015/slides/blsp/Rev_BIGDATA.pdf

[17] Jianqing Fant Yingying Fan+ and Yichao Wu§, High-Dimensional Classification, High-Dimensional Data Analysis: Volume 2 Frontiers of Statistics, https://doi.org/10.1142/7948 December 2010.

[18] M. Pavithra1 , and Dr. R.M.S.Parvathi , A Survey on Clustering High Dimensional Data

Techniques, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 11 (2017) pp. 2893-2899