International Journal of Scientific & Technical Development
(A Peer-Reviewed Bi-Annual Journal) ISSN: 2348-4047

A REVIEW OF DATA ANALYSIS TECHNIQUES AND DIMENSIONALITY REDUCTION IN HIGH DIMENSIONAL DATA MINING

***Pritika Mehra, Mini Singh Ahuja**
*Research Scholar GNDU Amritsar India
**Assistant Professor GNDU Regional Campus Gurdaspur India

Corresponding Author EmailID: pritikacsc.rsh@gndu.ac.in

Abstract
With the fast growth of computational biometric and e-commerce applications, high-dimensional data becomes very common. Thus, mining high-dimensional data is a critical problem of great practical importance. However, there are some unique challenges for mining data of high dimensions, including the curse of dimensionality and more crucial the meaningfulness of the similarity measure in the high dimension space etc. This paper provides a review of various challenges, techniques for analysis, and dimensionality reduction of high dimensional data.
Keywords: Dimensionality Reduction, High Dimensional data, principle component analysis, autoencoders

1. Introduction

The trend today is towards more observations, but even more so, to a very large number of variables—automatic, systematic collection of a large amount of detailed information about each observation. The observations could be curves, images or movies so that a single observation has dimension in the thousands or millions or billions while only tens or hundreds of observations for study. This is High Dimensional Data.High-dimensionality in combination with large datasets can be extremely challenging. High-dimensional data are relevant to a wide range of fields such as biometric, medicine, e-commerce, network security, and industrial applications. In order to use data characteristics, proper techniques and methods are needed to handle such high-dimensional data. Furthermore, data can have a typical characteristics and high-dimensional data structures, which means that conventional analysis techniques do not work well. To analyze extra useful information from high-dimensional data, novel approaches are required.

2. Challenges in High Dimensional Data

Among the many challenging issues concerning big data and high-dimensional data, we highlight the following five major challenges:

1. For high-dimensional datasets, there is the so-called curse of dimensionality: the searchable volume in the hyperspace becomes small, compared with the vast feasible search space. Thus, any solution procedure can only sample a subset of sparse points with essentially zero sampling volume in order to make sense of the vast datasets. Thus, it is a huge challenge with an almost impossible task for finding the global optimality. In addition, the distance measures required for problem formulations become less meaningful as any finite distance will result in an almost zero ratio between the distance measure and the vast distance needed to cover in the high-dimensional search space.
2. As the number of dimensions increases, the number of features also increases, often far more rapidly, which means that there is huge sparsity associated with such high-dimensional features. In addition, some correlation may exist between different dimensions, and thus features can be difficult to define.
3. For high-dimensional data, datasets tend to be unstructured, which may pose extra challenges to use. In addition, noise and uncertainties often exist in big datasets. Such noisy data can become more challenging to process and to apply any proper data mining techniques. For such problems, there is no analytical approach to provide insight even for a small subset of problems. Therefore, algorithms tend to be problem specific and even data specific. Thus, there is no generic approach in general.
4. As the number of dimensions increases, the possible combinations of clusters grow exponentially, and clustering becomes nondeterministic polynomial-time hard (NP-hard), and thus there are no efficient methods to deal with such challenging problems.
5. Even with the steady increase in speed of modern computers and the availability of cheaper parallel and cloud computing facilities, this does not ease the challenges of high-dimensional information

analysis. Efforts on developing new methods and tools are still highly needed. It may need a paradigm shift and a nonconventional way of thinking to problem-solving concerning high-dimensional data.

These challenges mean that new methods and alternative approaches are needed to solve such tough problems. In fact, heuristic and metaheuristic algorithms have been proven to be a promising set of alternative methods, especially those metaheuristic approaches based on nature-inspired optimization algorithms.

3. Methods of High Dimensional Data Analysis

a. Classification

It is a supervised learning technique. It arises frequently from bioinformatics such as disease classifications using high throughput data like micorarrays or SNPs and machine learning such as document classification and image recognition. It tries to learn a function from training data consisting of pairs of input features and categorical output. This function will be used to predict a class label of any valid input feature. Well known classification methods include (multiple) logistic regression, Fisher discriminant analysis, k-th-nearest-neighbor classifier, support vector machines, and many others. When the dimensionality of the input feature space is large, things become complicated.

b. Regression

In regression setting, one of the p variables is a quantitative response variable. Examples include, for instance in financial database, the variability of exchange rates today given recent exchange rates.

(i) Linear regression modeling:

Xi1 = a0 + a2Xi2 + • • • + apXip + Zi

Xi1 Response; Xi2, . . . ,Xip predictors;

(ii) Nonlinear regression modeling:

Xi1 = f(Xi2, . . . , Xip) + Zi .

(iii) Latent variable analysis:

In latent variable modeling, it is proposed that X = AS where X is a vector-valued observable, S is a vector of unobserved latent variables and A is linear transformation converting one into the other. It is hoped that a few underlying latent variables are responsible for essentially the structure we see in the array X and, by uncovering those variables, we get important insights.Principal Component Analysis (PCA) is an example. Here one takes the covariance matrix C of the observables X, obtains the eigenvectors, which will be orthogonal, places them as columns in an orthogonal matrix U and defines S = U'X.

Here we have the latent variable form with A = U. This technique is widely used for data analysis in sciences, engineering and commercial applications. The mathematical reason for this approach is that the projection on the space spanned by the first k eigenvectors of C gives the best rank k approximation to the vector X in a mean square sense.

c. Clustering

Here one seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar. Clustering high-dimensional data is the search for clusters and the space in which they exist. Thus, there are various kinds of clustering methods available in literature:

Subspace clustering

These approaches search for clusters existing in subspaces of the given high-dimensional data space, where a subspace is defined using a subset of attributes in the full space.

Subspace clustering methods will search for clusters in a particular projection of the data [12]. These methods can ignore irrelevant attributes and also problem is known as Correlation clustering. Two-way clustering, or Co-Clustering or Biclustering are known as the special case of axis-parallel subspaces. In these methods the objects are clustered simultaneously as the feature matrix consisting of data objects as they are span in rows and [11]. As in general subspace methods they usually do not work with arbitrary feature combinations. But this special case it deserves attention due to its applications in bioinformatics. CLIQUE-Clustering in Quest [13], is the fundamental algorithm used for numerical attributes for subspace clustering. It starts with a unit elementary rectangular cell in a subspace. If the densities exceeds the given threshold value, those cell are will be retained [5]. It applies a bottom-up approach for finding such units. First, it divides units into 1- dimensional equal units with equal-width bin intervals as grid. Threshold and bin intervals are the inputs for this algorithm [8]. It uses Apriori-Reasoning method as the step recursively from q-1-dimensional units to q-dimensional units using selfjoin of q-1. The total subspaces are sorted based on their coverage. The subspaces which are less covered are pruned. Based on MDL principle a cut point is selected and a cluster is defined as a set of connected dense units. A DNF expression that is associated with a finite set of maximal segments called regions is repre-

sented whose union is equal to a cluster [6].

Projected Clustering

Projected clustering tries to assign each point to a unique cluster, but the clusters may exist in different subspaces. The general approach uses a special distance function along with a regular clustering algorithm. PROCLUS-Projected Clustering, [2], associates with a subset of a low dimensional subspace S such that the projection of S into the subspace is a tight cluster. The pair (subset, Subspace) will represent a projected cluster. The number of clusters k and average subspace dimension n will be specified by the user as inputs [6]. It finds k-medoid in iterative manner and each medoid is associated with its subspace. A sample of data is used along with greedy hill-climbing approach and the Manhattan distance divides the subspace dimension. An additional data passes follow after the iterative stage is finished to refine clusters with subspaces associated with the medoids. ORCLUS-Oriented projected Cluster generation [3] is an extended algorithm of earlier proposed PROCLUS. It uses projected clustering on non-axes parallel subspaces of high dimensional space [9]

Hybrid Clustering

Sometimes it is observed that not all algorithms try to find a unique cluster for each point nor all clusters in all subspaces may have a result in between. It is because of having a number of possibly overlapping points [7]. The exhaustive sets of clusters are found necessarily. FIRES [4], can be used as a basic approach a subspace clustering algorithm. It uses a heuristic aggressive method to produce all subspace clusters [9].

Correlation Clustering

Correlation Clustering is associated with feature vector of correlations among attributes in a high dimensional space. These are assumed to persistent to guide the clustering process [2]. These correlations may found in different clusters with different values, and cannot be reduces to traditional uncorrelated clustering [6]. Correlations among attributes or subset of attributes results different spatial shapes of clusters. Hence, the local patterns are used to define their similarity between cluster objects [8]. The Correlation clustering can be considered as Biclustering as both are related very closely. In the biclustering, it will identify the groups of objects correlation in some of their attributes. The correlation is typical for the individual clusters [10].

Dimensionality reduction approaches try to construct a much lower-dimensional space and search for clusters in such a space. Often, a method may construct new dimensions by combining some dimensions from the original data.

4. Dimensionality Reduction

Dimension reduction is commonly defined as the process of mapping high-dimensional data to a lower-dimensional embedding. Applications of dimension reduction include, but are not limited to, filtering, compression, regression, classification, feature analysis, and visualization. Dimensionality reduction is a technique of reducing the feature space to obtain a stable and statistically sound machine learning model avoiding the Curse of dimensionality.

There are mainly two approaches to perform dimensionality reduction:

Feature Selection and Feature Transformation.

1. Feature Selection approach tries to subset important features and remove collinear or not-so-important features.
2. Feature Transformation also known as Feature Extraction tries to project the high-dimensional data into lower dimensions. Some Feature Transformation techniques are Principle Component Analysis (PCA), Autoencoders, Matrix Factorization, t-Sne, UMAP,etc

.

● Principle Component Analysis

Principle Component Analysis is an unsupervised technique where the original data is projected to the direction of high variance. These directions of high variance are orthogonal to each other resulting in very low or almost close to 0 correlations in the projected data.

● Autoencoders

Autoencoder is an unsupervised artificial neural network that compresses the data to lower dimension and then reconstructs the input back. Autoencoder finds the representation of the data in a lower dimension by focusing more on the important features getting rid of noise and redundancy. It's based on Encoder-Decoder architecture, where encoder encodes the high-dimensional data to lower-dimension and decoder takes the lower-dimensional data and tries to reconstruct the original high-dimensional data.

5. Conclusion

There is a dire need to focus on important issues of high

dimensionality problems and dimensionality reduction. High-performance computing approaches are best suitable for solving high dimensional data problems. The existing algorithms not always respond in an adequate same way when deal with extremely high dimension data due to exponential growth in the dimensionality and sample size.

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] GanGuojan, 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 NiponTheera-Umpon, Information Analysis of High-Dimensional Data and Applications, Mathematical Problems in Engineering(2015)

[15] [15]Wang 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] JianqingFantYingying 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.