Survey On Big Data DimensionalityReduction

Kanuparthi Prasasthi[1], Ponduru Likhitha[2], Venigandla Srilekha[3], Mohammad Rahamathunnisa[4], Kodukula Subrahmanyam[5], Nagamalleswary[6]
[1,2,3,4]*Research Student, Dept of CSE KL UNIVERSITY*
[5]*Prof at KL University, Dept of CSE*
[6]*Asst. Prof at KL University, Dept of CSE*

Abstract: **Dimensionality reduction is a common problem in scientific simulations, particularly in the field of transient simulations, where the number of variables can be very large. One approach to addressing this problem is to use diffusion maps, a nonlinear dimensionality reduction method that is based on the concept of diffusion on a graph. In transient simulations, the goal is often to understand the evolution of a system over time. This can be challenging when the system has a high number of variables, as it can be difficult to visualize and analyze the data. Dimensionality reduction techniques can be used for many numbers of variable detection, making it easier to analyze and understandthe data.**

Diffusion maps is a nonlinear dimensionality method coming from the idea of diffusion on a graph. It uses the eigenvectors of the graph Laplacian to identify useful and needed variables in the system and to project the given data onto a dimensional space which is lower . This allows for visualization and also analysis of the data in a more manageable form. In the context of transient simulations, diffusion maps can be used to identify the most important variables and to track the evolution of the system over time. It can also be used to identify patterns and relationships in the data that may not be evident in the original, high-dimensional space. Overall, the use of diffusion maps in transient simulations can provide valuable insights into the behavior of the system and can facilitate the understanding and analysis of complex, high-dimensional data.

Index Terms: **Big data, Reduction, data, high-dimensional, low-dimensional, records, dimensionality.**

INTRODUCTION

Data Reduction: Data reduction is a common techniqueused in data analysis and data management to simplify and summarize large amounts of data in a more manageable form. It involves reducing the size of the data set by eliminating unnecessary or redundant data, and often involves aggregating the data into more general categories or summary statistics. This can be help in pattern identification for the data, as well as for visualizing and communicating the results of an analysis.[1]

There are several different methods that can be used for data reduction, including sampling, aggregation, and dimensionality reduction. Sampling involves selecting a representative subset of the data to analyse, rather than analysing the entire data set. Aggregation involves combining similar data points or records into larger groups and calculating summary statistics such as or irrelevant features or any kind variables that are present in data collection to make data easier to analyse and interpret.[2]

Data reduction can be particularly useful when while using data sets which are approximately large, it can be useful to reduce the time and resources required to analyse the data. It can also be useful for identifying trends and data which contain patterns which temporarily cannot be immediately apparent when looking at the data in its raw form. However, it is important to carefully consider the methods and techniques used for data reduction, as they can have a significant impact on the results and conclusions of an analysis.[3]

Yes, that's correct. In many cases, the data that we work with has already been transformed from its original form in some way. For example, analog data (such as measurements taken with a physical instrument) may needto be converted to digital form to be analysed or stored on a computer. Similarly, data that is already in digital form may mean or median for each group. Dimensionality reduction involves identifying and removing redundant.

Figure1: Data Reduction Process

Yes, the example of the Kepler satellite illustrates the importance of data reduction in situations where the amount of data being collected is simply too large to be transmitted or analysed in its raw form. By applying techniques such as co-adding, pre-selection, and only processing relevant pixels, the data reduction process is able to significantly minimize the data set size untilmaintaining the key information that is needed for analysis.

The use of data reduction in wearable devices for health monitoring and diagnosis is another area where data reduction can be particularly useful. In these types of applications, it is often important to reduce the transmitted data amount for conserving battery life and reduce the burden on the device's processor. By carefully selecting and transmitting require some editing or cleaning to remove any errors or inconsistencies.

Smoothing and interpolation can also be important techniques in data reduction, especially when working with data that is noisy or has a lot of variation. Smoothing involves applying a mathematical function to the data to smooth out any sudden changes or spikes and can help to reveal underlying trends or patterns that might not be apparent in the raw data.[2][3] Interpolation, on the other hand, involves estimating the value of a variable at a specific point based on its known values at other points. This can be useful when working with data that is sampled at irregular intervals, or when trying to fill in missing data points [4].

It is uniquely identified in grasping the errors in the data when performing data reduction. This can help to identify any potential biases or uncertainties in the data and can inform the choice of methods and techniques used to reduce and analyse the data. only the data that is relevant for diagnosis, data reduction can help to extend the battery life of the device and improve its overall performance.

Dimensionality Reduction:

Dimensionality reduction is a technique used to minimize the features and dimensions count for dataset. It is often used when dealing with datasets that have a high numberof features, as it can make it easier to visualize the data and build models that perform well on it.

Dimensionality reduction is a technique used in decreasein the variable dimensions number in the data collection while collecting more information or matter as much we need . It is often used to pre-process data before building a machine learning model, as it can improve the model's performance and reduce the risk of overfitting.[5]

There are two important main types in reduction of dimensionality: Feature Selection and Extraction . Selection includes subset selection of features which are original and later on Extraction includes transforming theoriginal features into the set which is new features that are included with the features which are original .

Some common techniques for reduction in dimensionality include principal component analysis (PCA), linear discriminant analysis (LDA), and singular value decomposition (SVD). These techniques can be applied to both continuous and categorical data, and they are also used in different kinds of contexts, includes recognition in image , NLP, predictive modelling.

Overall, dimensionality reduction is a useful tool for improving the performance and interpretability of machine learning models, and it can help to reduce the risk of overfitting and improve the generalization of the model to new data.[1]

Techniques:

Data reduction techniques are techniques used to compress or summarize large data sets to make them more manageable or easier to analyze. These processes can be useful in minimizing the size of data sets, increasethe speed of data processing, or simplify data analysis tasks. Here are some examples of data reduction techniques:

Sampling: Sampling involves selecting a subset of the data to represent the entire data set. This can be useful when the data set is too large to be analyzed in its entirety.

Sampling can be a useful technique fordimensionality

reduction because it allows you to select a representative subset of the data that captures the main characteristics of the entire dataset. This can be particularly useful if the data is very large, and it would be impractical to work with all of the data at once. By sampling the data, you can select a smaller, more manageable dataset that still contains enough informationto be meaningful. [5]
There are different methods for sampling data for dimensionality reduction, including random sampling,stratified sampling, and cluster sampling. Selection of sampling process will be dependent on specificcharacteristics of the data and the goals of the analysis. Ingeneral, it is important to ensure that the sampling methodis appropriate for the data and that the sample is representative of the entire dataset.

Aggregation: Aggregation involves combining multiple data points into a single summary statistic. For example, you might aggregate a large dataset by calculating the average value for each variable. In dimensionality reduction, aggregation refers to the process of combining or merging multiple data points or samples into a single representative value. This is often done to reduce the complexity of a dataset or to simplifythe data for further analysis or modelling.
There are many different techniques for aggregation in dimensionality reduction, including mean aggregation, median aggregation, and mode aggregation.Mean aggregation involves considering the value which is average with all the points of that data while median aggregation involves finding the middle value in the data set. Mode aggregation involves finding most commonly occurred value from the data collection . Other techniques for aggregation in dimensionality reduction include min- max normalization, which scales the data between a minimum and maximum value, and standardization, which scales the data to have a zero as mean and one as SD.
Aggregation can be useful in dimensionality reduction because it can help to reduce the noise or random variability in the data, and it can also help to highlight trends or patterns that are included in data, and which are not approximate when looking at individual data points. However, it is important to carefully consider the appropriate aggregation technique to use, as different techniques may lead to different results and may beappropriate depending on the specific data functions andto achieve the main part of analysis.

Dimensionality reduction: Dimensionality reduction techniques helps in decrease in variables number or rulesin a data set. This is useful when the data set has many irrelevant or redundant features, which can make it moredifficult to analyse.[6]

Data compression: Data compression techniques reduce in datasets size through encoding particular data in a relevant compact format. This can be useful when the data needs to be transmitted or stored efficiently.

Feature selection: Feature selection involves identifyingthe most important features in a data set and discarding therest. This can be useful when the data set has manyirrelevant or redundant features.

Data transformation: Data transformation techniques modify the data in some way to make it more suitable for analysis. For example, you might transform a dataset by scaling the data or applying a mathematical function to the data.

Examples of Dimensionality Reduction:
Here are some examples of dimensionality reduction techniques:

Principal Component Analysis (PCA): This reminds ina technique of dimensionality reduction which is linear that arranges the given data in a dimensionality space which is lower by identifying all directions of variance which is maximum in the data.

Singular Value Decomposition (SVD: This reminds the factorisation matrix technique that is useful in reduction of dimensionality . It decomposes a matrix into the product of three matrices, which can then be useful in representing the data in a Dimensional space which is lower

T-distributed Stochastic Neighbor Embedding (tSNE): This is a technique in dimensionality from the non – linear reduction that can collab the data to a dimensionality space which is lower while creating the structure and relation between the data points.

Autoencoders: These are neural network models that aretrained to reconstruct the input data from a lower-dimensional representation, or "latent space." They can be used for dimensionality reduction by training the model on the input data and then using the latent space asthe reduced dimensional representation of the data.

Feature Selection: This reminds in selecting a subset process of the most useful and understandable features from all the data collections to use in all the machine learning model. This can be done using techniques such as recursive feature elimination or feature importance measures.

To calculate the dimensionality reduction, you can use one of these techniques on your dataset and specify the desired number of dimensions or the amount of varianceyou want to retain. The output of the dimensionality reduction will be a transformed version of the original data with fewer dimensions.

Challenges of Dimensionality Reduction:

Information loss: One of the main challenges of dimensionality reduction is that it often involves a trade-off between the number of dimensions and the amount of information retained. As the number of dimensions is reduced, some information is inevitably lost.

Curse of dimensionality: The curse of dimensionality refers to the difficulty of working with high-dimensionaldatasets, which can cause many machine learning algorithms to perform poorly. Dimensionality reduction can help mitigate this problem by reducing the number ofdimensions

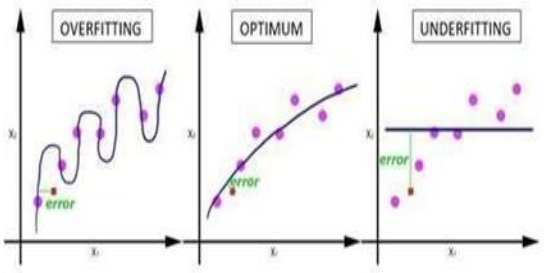

Figure2: variations in dimensionality[2]

Overfitting: When using dimensionality reduction techniques, it is important to avoid overfitting the data. This can occur if the dimensionality reduction method tries too hard to fit the training data, leading to poor generalization to unseen data.

Choosing the right method: There are many different techniques for performing dimensionality reduction, and each has its own strengths and weaknesses. Choosing theright method for a particular dataset can be challenging, and it may be necessary to try several different methods to find the one that works best.

There are several ways to calculate overfitting. One common method is to use the difference between the performance of the model on the training dataset and theperformance on a validation dataset. If the model performs significantly better on the training dataset compared to the validation dataset, it is likely to be overfitting. Another method is to use cross validation, which involves dividing the training dataset into multiplesmaller datasets, training the model on each of the smallerdatasets, and evaluating its performance on the remaining data. Accurate estimation can be provided more for the model's generalization performance and help identifyoverfitting.

It's also possible to use techniques such as regularization, which involves adding constraints to the model help in decrease the parameter number and also to prevent overfitting, early stopping, which involves training the model until the performance on the correction of datasetstarts to get down and then stopping the particular training to prevent overfitting.

Complexity: Some dimensionality reduction techniques, such as kernel PCA, can be computationally intensive andmay not be suitable for large datasets.

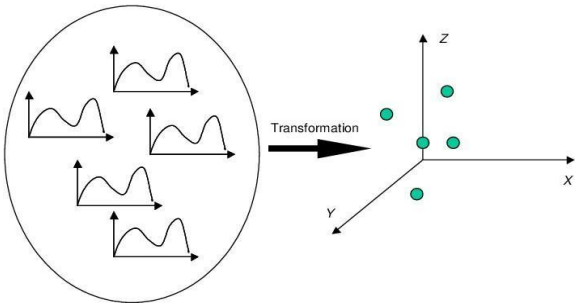

Figure3: Transferring of data points[3]

Difficult in interpretation: The data presentation for thereduction in dimensionality difficult to interpret, making it harder to understand the relationships between the features and the target variable.

LITERATURE REVIEW

1.Laurens van der Maaten, Eric Portman, H. Jaap van den Herik. Dimensionality Reduction: A Comparative Review. Published on January 2007 in ResearchGate

Dimensionality reduction is a technique used to decreasein some of features or any for the dataset while retainingas much information as possible. It is often used inmachine learning and data analysis to improve the performance of models, decrease in training cost and evaluating designs , and improve the interpretability of the results. This includes many different ways in dimensionality reduction, each and every thing with its own merits and demerits , and choosing the appropriate method depends on the data features and involves in achievement of goals in analysis.

The most widely used methods for reduction of dimensionality is PCA. The technique is a dimensionality reduction which is linear which identifiesthe directions of the given data and also having the variance which is highest and arranges the data in a dimensionality space which is lower along those directions. PCA is generally used in decrease indimensions of huge datasets and to visualize the structureof the data.

Another popular method for dimensionality reduction is singular value decomposition (SVD). SVD isa technique which is used under factorisation technique that the matrix decomposes into the three matrices product: the singular matrix on the left, a diagonal matrix of singular values, and a right singular matrix. SVD can be used to identify the underlying factors or components that explain data variance and is helpful to arrange the particular data on to a dimensional space which is lower . Other methods for dimensionality reduction include independent component analysis (ICA), which helps to decompose the data collection into the group of statistically independent components, and non-negative matrix factorization (NMF), which with the product of two matrices which are non-negative and helps in decomposing a matrix.

There are also many techniques for non linear dimensionality reduction, such as t distributed stochastic neighbor embedding (t-SNE) and multidimensionalscaling (MDS). These techniques can be particularly needed in visualizing high-dimensional datasets and for uncovering non linear structure in the data.

In addition to these methods, there are also various techniques for feature selection, which is the goalof subset selection of the very properly relevant features from a dataset, such as random forests, mutual information, and lasso regression. Feature selection can be used in conjunction with dimensionality reduction to further improve the interpretability and performance of machine learning models.

While, passing through these models we have observed a lot in all of the methods that are in the Dimensionality reduction

2.Dr. S. Vijaya Rani, Ms. S. Maria Sylviaa.DIMENSIONALITY REDUCTION - A STUDY.Published Online April - May 2016 in IJEAST

Dimensional reduction is a technique that lead to decrease the dimensions in number or characters in a data collection in finding most of the important information as possible. This can be useful when working with high- dimensional datasets, as it can make it easier to visualizeand analyze the data and can also improve the performance of machine learning algorithms.

There are several techniques for dimensional reduction, including different types such as Principle and independent component analysis PCA and ICA , Singular Value Decomposition (SVD) , These techniqueswork by identifying patterns in entire data and showing the given data in the dimensionality space which is lower, which can help to identify trends and relationships that may not be apparent in the original data. Dimensional reduction can be useful in different kinds of applications which include the recognition of images , NLP, financialobservations. It can also be used to identify underlying patterns and trends in complex datasets and can be provedbetter to perform under algorithms of machine learning.

Overall, dimensional reduction is a powerful tool for analyzing and understanding complex datasets and can be a valuable tool for data scientists and analysts working in a variety of fields.

METHODOLOGY

We conducted a survey on big data dimension reduction. We worked on the survey by gathering responses from ITindustry employees via a Google

form. By giving their input on the form. Through the responses, we calculated the accuracy rate for dimensionality reduction usage in the IT industry and the scope of a future study to implement it to get a better accuracy rate than the presentaccuracy rate on dimensionality reduction.

There are different types of methods to go further in the process. In the process of finding the possible way. Thetechniques which are actually included in dimensionalityreduction are

1.Sampling

2.Aggregation

Some techniques are named and based on the survey done on the dimensionality reduction The further process can be based on the technique named Principal Component Analysis (PCA).

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that can be used to project high-dimensional data onto a lower-dimensional space by finding the directions of maximum variance in the data. Here is an outline of the steps involved inperforming PCA:

Standardize the data: PCA is sensitive to the scale of thevariables, so it is important to standardize the data to havezero mean and unit variance before applying PCA.

Compute the covariance matrix: The covariance matrix isa measure of the variability of the data. It is a square matrix with dimensions equal to the number of variables, and the i,j-th element is the covariance between the i-th and j-th variables.

Compute the eigenvectors and eigenvalues of the covariance matrix: Eigenvectors are vectors that do not change direction when a linear transformation is applied to them. The eigenvectors of the covariance matrix are the directions of maximum variance in the data. The eigenvalues are the magnitudes of the variance along thecorresponding eigenvectors.

Select the number of components: To determine the number of components to retain, we can either specify thenumber of components we want to keep or we can select a cutoff for the explained variance. For example, we can choose to retain the top 2 components, which capture themost variance in the data.

Transform the data onto the new subspace: To transform the data onto the new subspace, we multiply the standardized data by the matrix of eigenvectors. This willresult in a new matrix with the same number of rows as the original data, but with a reduced number of columns.

RESULTS & OUTPUT

The primary purpose of this research is to conduct a survey on big data dimensionality reduction. To know how dimensionality reduction is useful in the IT industry and collect their responses. Dimensionality reduction involves several techniques, some of which are: 1) analysis of major components (PCA); 2) Decomposition of singular value (SVD) 3) Stochastic integration brokendown into T (tSNE).

In the IT industry, principal component analysis (PCA) was the most commonly used technique for dimensionality reduction. 50–75% improves data accuracy in dimensionality reduction. Dimensionality reduction is used for several dataset types, like voice, image, and data. While using the maximum dimensionality reduction, we don't lose any data or information from the dataset.

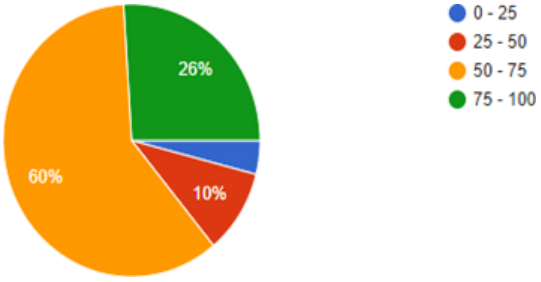

Figure4: Data accuracy

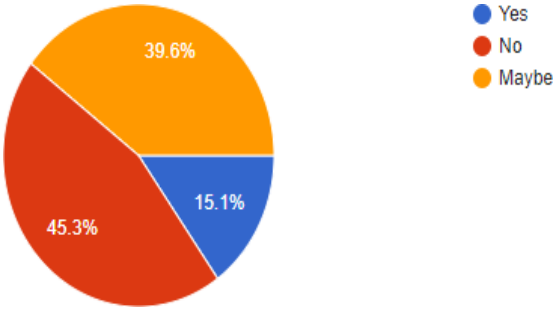

Figure5: Data loss

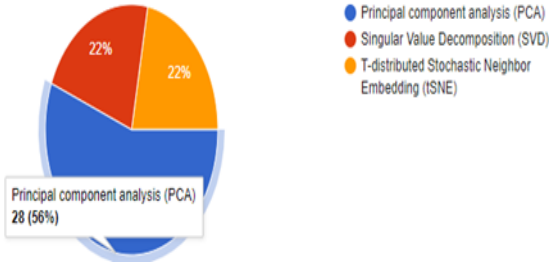

Figure6: Dimensionality Reduction Techniques

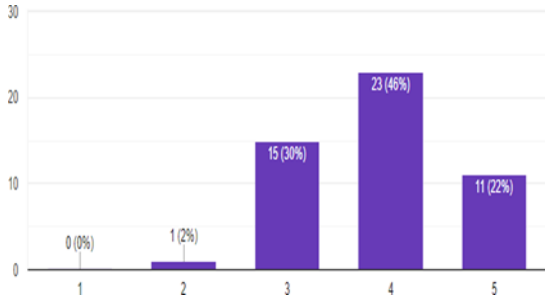

Figure7: Dimensionality Reduction Accuracy

There are several techniques to reduce a large dataset. Inthe dimension reduction technique, we convert the higher-dimensional dataset into a lesser-dimensional dataset. According to our survey, principal component analysis (PCA) accounted for 56%, singular value decomposition (SVD) for 22%, and t-distributed stochastic neighbor embedding (t-SNE) for 22%. PCA has more accuracy than the IT industry. In the future, we will improve our technology for dataset reduction with better than present accuracy without losing any data or information from the data set.

CONCLUSION

Dimensional reduction is a technique in physics and mathematics that involves simplifying a problem by reducing the number of dimensions that it is defined in. This can be done for a variety of reasons, such as to make a problem more tractable, to eliminate redundancies in theproblem, or to reveal underlying structures or symmetries.There are several different approaches to dimensional reduction, including Kaluza-Klein theory, spontaneous symmetry breaking, and the renormalization group.

One of the key ideas behind dimensional reduction is that,in many cases, the behavior of a system can be better understood by studying a lower-dimensional projection ofthe system rather than the system in its full, high- dimensional space. This can be especially useful when the high-dimensional space is too complex or difficult to work with directly. However, it is important to note that dimensional reduction is not always possible, and even when it is, it is not possible to collect all the important information in the reduced system.

REFERENCE

[1] Laurens van der Maaten, Eric Portman, H. Jaap Van Den Herik. Dimensionality Reduction: A Comparative Review. Published in January 2007 in ResearchGate

[2] Dr. S. Vijaya Rani, Ms. S. Maria Sylviaa. DIMENSIONALITY REDUCTION - A STUDY. Published Online April - May 2016 in IJEAST

[3] Weikuan Jia, Meili Sun, Jian Lian & Sujuan Hou. Feature dimensionality reduction: a review. Published: 21 January 2022, SpringerLink

[4] S. Velliangiria, S. Alagumuthukrishnanb, IwinThankumar josephc.A. A Review of Dimensionality Reduction Techniques for Efficient Computation.Published:27 February 2020, ScienceDirect

[5] Raji Ramachandran, Gopika Ravichandran, Aswathi Raveendran. Evaluation of Dimensionality Reduction Techniques for Big data. Published:23 April 2020, IEEE

[6] Soudagar Londhe. Dimensional Reduction Techniques for Huge Volume of Data. Publishedon:2022-03-01, Ijraset

[7] Maha Alkhayrat, Mohamad Aljnidi & Kadan Aljoumaa.A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. Published: 03 February 2020, SpingerOpen

[8] C.O.S. Sorzano, J. Vargas, A. Pascual Montano. A survey of dimensionality reduction techniques. Published on12 Mar 2014, arxiv

[9] Ali Ghodsi. Dimensionality Reduction: A Short Tutorial. Published on:2006, Clemson.

[10] G. Thippa Reddy, M.Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, Thar Baker. Analysis of Dimensionality Reduction Techniques on Big Data. Published on:2020, IEEE