IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REDUCING DIMENSIONALITY IN TIME SERIES DATA USING NEURAL AUTO ENCODER TECHNIQUE

[1]Mrs. M. P. Rekha, [2]Dr. K. Perumal,

[1]Research Scholar, [2]Professor

[1]Department of Computer Applications, School of Information Technology

[1]Madurai Kamaraj University, Madurai, Tamil Nadu, India

Abstract – The most important problem in time series data is High Dimensionality of data. The high dimensionality of data has more number of attributes under considerations. The process of dimensionality reduction is to reduce the number of random variables or attributes. The various techniques used to reduce the dimensionality of data. The Dimensionality reduction techniques have feature extraction and feature selection process. This paper focuses the feature extraction using Neural Auto Encoder technique is used to reduce the dimensionality of data. Compare with other techniques, the proposed Neural Auto Encoder technique produce the high level of reducing the dimensionality of data. The proposed method gives more accuracy in time series data.

Index Terms **- Data mining, Time series data, Dimensionality reduction technique, Feature Extraction**.

I. INTRODUCTION

1.1 Time Series Data

A time series is the series of elements in time request. A time series is an arrangement of progressive equivalent stretch moments. It investigates time series information to extricate significant data and different qualities of information. Time-series information investigation turns out to be vital in such countless businesses like monetary enterprises, drugs, online media organizations, web specialist co-ops and some more.

1.2 Time Series Analysis

The Series of information focuses recorded throughout a predetermined timeframe is called Time-series information. One of the significant goals of the examination is to figure future worth. Extrapolation is involved when anticipating with the time series examination which is very perplexing. Time series examination can be valuable in after.

1.2.1 Pattern: Increasing or diminishing example has been seen throughout some undefined time frame. For this situation, the continuously expanding fundamental pattern is noticed. For example the count of travelers has expanded throughout some undefined time frame.

1.2.2 Irregularity:

Refers to cyclic example. A comparable example that rehashes later a specific time frame.

1.2.3 Heteroscedasticity:

Refers to Non-steady fluctuation or differing avoidance from the mean throughout some stretch of time.

1.3 Time Series Data Mining

The Time Series Data Mining have the accompanying angles:

1.3.1 Ordering (Query by Content):

Inquiry by content in time series data sets has arisen as a space of dynamic interest. This additionally incorporates an arrangement matching assignment which has for some time been separated into two classes: entire coordinating and aftereffect coordinating. Entire Matching is a question time series is matched against a data set of individual time series to distinguish the ones like the inquiry. Aftereffect Matching is a short inquiry aftereffect time series is matched against longer time series by sliding it along the more extended succession, searching for the best matching area.

1.3.2 Clustering :

Clustering is like arrangement that orders information into gatherings; how-ever, these gatherings are not predefined, yet rather characterized by the actual information, in light of the likeness between time series. It is frequently alluded to as unaided learning. The grouping is typically refined by deciding the similitude among the information on predefined ascribes.

1.3.3Prediction (Forecasting):

Prediction can be considered a kind of bunching or characterization. The thing that matters is that forecast is foreseeing a future state, rather than a current one. Many time series forecast applications can be seen in financial spaces, where an expectation calculation normally includes relapse investigation

1.3.4 Summarization

A summarization of the information is helpful and essential. A measurement synopsis of the information, for example, the mean or other factual properties can be effectively registered. Outline can likewise be seen as a unique kind of bunching issue that maps information into subsets with related straightforward (text or graphical) portrayals and gives a more elevated level perspective on the information.. The synopsis might be done at various granularities and for various aspects.

1.3.5 Inconsistency Detection

In time series information mining and observing, the issue of distinguishing anoma-lous/astonishing/novel examples has drawn in much consideration. Rather than aftereffect coordinating, irregularity discovery is recognizable proof of already obscure examples.. From an overall perspective, a bizarre conduct is one that goes astray from "typical" conduct.

1.4 Dimensionality

Dimensionality in insights alludes to the number of characteristics a dataset has. For instance, medical services information is infamous for having huge measures of factors (for example pulse, weight, cholesterol level). In an ideal world, this information could be addressed in an accounting page, with one section addressing each aspect. Practically speaking, this is hard to do, partially on the grounds that numerous factors are between related (like weight and pulse).

1.5 High Dimensionality

High Dimensional implies that the quantities of aspects are incredibly high — so high that computations become amazingly troublesome. With high layered information, the quantity of elements can surpass the quantity of perceptions. For instance, microarrays, which measure quality articulation, can contain many examples. Each example can contain a huge number of qualities.

1.6 Dimensionality Reduction in Time Series

Dimensionality Reduction is the method involved with lessening the quantity of aspects in the information either by barring less helpful highlights (Feature Selection) or change the information into lower aspects (Feature Extraction). Dimensionality decrease forestalls over fitting. Over fitting is a peculiarity where the model gains excessively well from the preparation dataset and neglects to sum up well for inconspicuous genuine information.

Dimensionality reduction strategy can be characterized as, "It is a method of changing over the higher aspects dataset into lesser aspects dataset guaranteeing that it gives comparative information.It is normally utilized in the fields that arrangement with high-layered information, like discourse acknowledgment, signal handling, bioinformatics, and so on It can likewise be utilized for information representation, commotion decrease, bunch investigation, and so forth

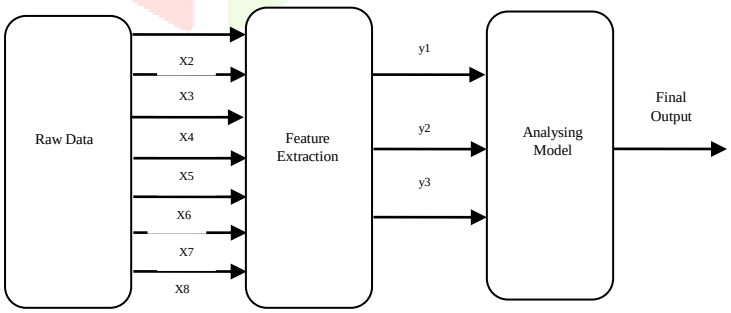

Fig.1 Feature Extraction Process

1.7 Methods Of Dimensionality Reduction

Dimensionality decrease utilizing the accompanying methods, Feature Selection and Feature Extraction. Highlight choice depends on excluding the excess and unimportant elements are disregarded. Highlight extraction considers the entire data content and guides the valuable data content into a lower layered element space. Include extraction is for making a new, more modest arrangement of highlights that actually catches the greater part of the helpful data.

Three kinds of Feature Selection for Dimensionality Reduction are Recursive Feature Elimination, Genetic Feature Selection and Sequential Forward Selection. Kinds of Feature Extraction for Dimensionality Reduction, AutoEncoders, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA).

II. LITERATURE SURVEY

In the work **Michael Steinbach and Levent Ertöz**(2014)was discussed The Challenges of Clustering High Dimensional Data is to find data compression. They provide a short introduction to cluster analysis, focus on challenge of clustering high dimensional data. They present a overview of several recent techniques, include a more detailed description of concept based clustering approach.

In **Andrew McCallum and Kamal Nigam(2014)** discussed about **Efficient Clustering of High-Dimensional Data Sets.** They presented a new technique for clustering large high dimensional datasets. The key idea involves an approximate distance measure to divide the data into subsets called canopies. Then cluster is performed by measure exact distances only between points that occur in common canopy. Using the cheap distance metric , the reduction in computational cost comes without any loss in clustering accuracy.

Chao Chen and **Novi Quadrianto(2016)** was discussed about **Clustering High Dimensional Categorical Data**. Analysis of categorical data is a challenging task. In that paper, they compute topographical features of high-dimensional categorical data. They are explained an algorithm to extract modes of the underlying distribution effectively. These features provide a geometric view of data and applied to visualization and clustering of real world datasets.

In **Guha S., Rastogi R., Shim(2012)**was discussed about The rapid growth of analysing high dimensional data in various new application domains, like bioinformatics and e-commerce. Many organizations have massive amounts of data containing valuable information for running and building a decision making system.

In the work **J. Han and M. Kamber(2010)** This work study and to analyse high dimensional and large amount of data for effective decision making. Researchers and practitioners are very eager to analyse these datasets.

In **A. Jain, M. N. Murty(2011)**discussed about before analyse the data mining models, the researcher will analyse the challenges of attribute selection, the curse of dimensionality, reduce redundancy, data labelling and the specification of similarity in high dimensional space for analysing high dimensional data set .

In **Zhang T., Ramakrishnan R(2012)** discussed In data mining, the objects have hundreds of attributes or dimensions. Clustering in high dimensional data spaces presents a tremendous difficulty, much more than in predictive learning .

III. PROBLEM DEFINITION

Dimensionality reduction is utilized for scaling back input information is more applicable for additional investigation. Diminished dataset jelly change from bigger dataset and with no deficiency of significant highlights. It will turn out to be not difficult to identify and use from genuine information. Examination of Factor Analysis, Principal Component Analysis and wavelet investigation are applied to stable burning and found that incessant trait of comparative subjectively and quantitatively. However, the unsound and transient ignition is applied, Factor Analysis and Principal Component Analysis are not achievable. Factor Analysis and Principal Component Analysis are have limit and work effectively with consistent and organized peculiarities. Head Component Analysis and Factor Analysis requires more computational time and memory utilization since it requires entire preparing information to remove vectors.

IV. PROPOSED IMPLEMENTATION

4.1 Neural Auto Encoder Technique

Neural Auto Encoder is an unsupervised Artificial Neural Network .It encodes the information by packing it into the lower aspects and afterward translates the information to reproduce the first info. Auto Encoder has two sections specifically Encoder and Decoder. Encoder has input information and pack it. The encoder part eliminates all clamor and pointless data. The result of Encoder is called Bottleneck space.

In the proposed Auto Encoder technique, the feed information should scaled between 0 and 1 utilizing min max scalar. The dimensionality decrease will be remove the jug neck layer and it is use to diminish the aspects. This cycle is called Feature Extraction.

In the proposed strategy, the encoder part assists with learning significant secret highlights present in the information, to diminish the remaking mistake. Another arrangement of mixes of unique highlights is generated.The proposed technique for Auto Encoder is Deep Auto Encoder. In this technique, the encoder and decoder are even.

The Auto Encoder system has the info layer (X1 , X2 , . . . , Xn), stowed away layer (H1 , H2 , . Hm), and result layer (Y1 , Y2 , . . . , Yn) and the loads of stowed away layer represent properties of the info signal.

Φ : X-> F

Ψ: F-> X

Φ, Ψ= argnmin || X-(Ψ 0 Φ) X||2

Φ, Ψ

Φ – Encoder work, X – Original Data, F – Latent Space present at the bottleneck, ψ means the decoder work.

The result, for this situation, is as old as info work. The encoding organization can be addressed by the standard neural organization work went through an enactment work, where z is the idle aspect.

Likewise, the unraveling organization can be addressed in a similar manner, however with various weight, inclination, and possibly enactment capacities being utilized. The autoencoders convert the contribution to a diminished portrayal which is put away in the center layer called code. The data from the info has been compacted and by removing this layer from the model, every hub would now be able to be treated as a variable.

V. EXPERIMENTAL SETUP

5.1 Sample Data Set

Using the sample dataset the proposed techniques used to reduce the dimensionality of data. The Encoder part in AutoEncoder technique is to compress the data. It removes the noise and unhelpful information

Table 1Sample Data Set

Date	Time	Visibility	Temperature C	Temperature F
13/03/2012	12:00	10	-1.1	26
13/03/2012	23:00	10	-1.1	26
14/03/2012	21:15	6	-1.7	26
15/03/2012	8:45	0.75	-2.2	26
15/03/2012	19:15	10	-3.3	24
16/03/2012	7:00	10	-3.9	22
17/03/2012	15:45	5	-5	21
18/03/2012	3:15	10	-6.1	18
18/03/2012	14:00	9	-6.1	18
19/03/2012	1:15	10	-6.1	18
19/03/2012	12:15	10	-6.1	18
19/03/2012	23:15	10	-6.1	18
20/03/2012	10:30	10	-5.6	18
20/03/2012	21:30	10	-5.6	19
21/03/2012	9:00	10	-6.7	17
21/03/2012	19:30	6	-7.8	16
22/03/2012	7:00	5	-8.9	14
22/03/2012	17:30	0.5	-9.4	13
23/03/2012	5:00	1	-10	12

VI. RESULTS AND DISCUSSIONS

The proposed technique was tried with climate estimating datasets. The proposed framework was straightforward to comprehend and it is executed with the assistance of python and run on work area Pc with 3.60 GHz Intel and 4 GB RAM and 1 TB HDD. The result is more precise and along these lines it can fundamentally work on the presentation of client.

In weather dataset, the 24 number of sections are set. Each data has high dimensionality ie., every data has many attributes. Using the proposed Neural Auto Encoder technique the dimensionality has been diminished. The data has unassuming number of qualities. The 24 sections has been diminished. Directly following decreasing the dimensionality the component extraction strategy is used to isolate the part of data.

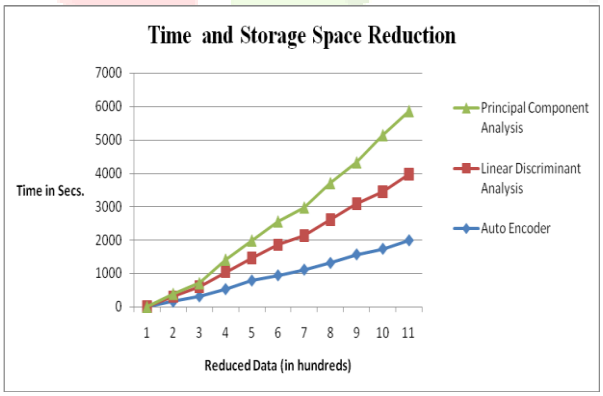

Fig 2. Time and Storage Space Reduction

In Figure 2 Shows the comparison of time and storage space reduction with existing systems and proposed system. Compared with existing Principal Component Analysis and Linear Discriminant Analysis, the proposed Neural Auto Encoder technique reached the highest accuracy and dimension reduction with low execution time.

Table 2 Comparison of Proposed Method vs Existing Method

Proposed vs Existing Method	Accuracy (%)	Execution Time (%)	Dimension Reduction(%)
Principal Component Analysis	84.57	73.68	84.57
Linear Discriminant Analysis	87.30	62.56	85.35
Proposed Neural Auto Encoder Method	92.45	61.39	91.78

VII. CONCLUSION

The proposed work reduces the dimensionality of information utilizing the Neural Auto Encoder procedure. Contrast and a current procedure, the proposed Neural Auto Encoder technique produce the significant degree of diminishing the dimensionality of information. The proposed technique gives more accuracy in time series information.

REFERENCES

[1] Liyanaarachchilekamalgechamara, Yan yang, Guang Bin, "Dimensions Reduction with Extreme Learning Machine", IEEE Transactions, vol 25, no 8, Aug 2017.

[2] Yanni Dong, Bo du, Liangpeizhang, "Dimensionality reduction and classification of Hyperspectral Images using Ensemble Discriminative Local Metric Learning", IEEE Transaction, vol 55, no 5 May 2017.

[3] . Priyanka jandal, Dharmaenderkumar, "A review on Dimensionlityreduction techniques", International Journal of Computer Applications,
vol 173, no 2 sep 2017.

[4] KHALID RAZA Centre for Theoretical Physics, JamiaMilliaIslamia, New Delhi-110025, India‖ APPLICATION OF DATA MINING IN BIOINFORMATICS‖ Khalid Raza / Indian Journal of Computer Science and Engineering Vol 1 No 2,114-118

[5] Wendy Foslien, Valerie Guralnik, Karen Zita Haigh Honeywell Laboratories, 3660 Technology Drive, Minneapolis, MN 55418‖ Data Mining For Space Applications‖ SpaceOps 2004 -Conference

[6] Data Mining Concepts and Techniques – Jiawei Han &MichelineKamber

[**7]** Nan Jiang and Le Gruenwald The University of Oklahoma, school of Computer Science, Norman, OK 73019, USA

[8] Dr. MohdMaqsood Ali1 Asst. Professor and Head of Marketing Department, Jazan Community College, Jazan University, Jazan Kingdom of Saudi Arabia 1 maqsoodphd@gmail.com" Role Of Data Mining In Education Sector" International Journal Of Computer Science And MobileComputing

[9] G. E. Hinton and R. S. Zemel, ''Autoencoders, minimum descriptionlength and Helmholtz free energy,'' in Proc. Int. Conf. NeuralInformation Process. System, 1993, pp. 3–10.

[10]. T. Ahmad, R. A. Fairuz, F. Zakaria, and H. Lsa, ''Selection of a subsetof EEG channels of epileptic patient during seizure using PCA,'' inProceedings. World Science Engineering Acad. Social Science.(WSEAS), 2008, pp. 270–273.

[11] Pouria fewsee, fakhrikarray," Dimensionality Reduction foremotional speech recognition", International conference on privacy,security and trust 2012.

[12] Tsuge S, Shishibori, Kuraiwa, "Dimensionality reduction using NMFfor information retrieval", IEEE international conference on systemsand cybernetics 2001.

[13] Dimension Reduction for Individuality of Handwriting in Writer Verification" , IEEE International conference on intelligent system
design and applications, 2013.

[14] Xinwei Jiang, Junbin Gao, Tiangiang Wang, Daming Shi, " ADimensionality Reduction Algorithm based on thin plate splines",IEEE Transaction on Cybernetics vol 44, no 10 , OCT 2014.

[15] Arul kumar, Elavarasanm,"A survey on Dimensionality reductiontechnique", IJETTCS, vol 3 issue 6, Dec 2014.

[16] Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE "Survey of Clustering Algorithms‖ IEEE Transactions OnNeural Networks, Vol. 16,No.3, May 2005.

[17] Vikas Gupta, Prof. Devanand‖ A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues‖ International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 1ISSN2229-5518.

[18]. Prof. K. Vijayalakshmi M.C.A., M.Phil., — Survey Of Data MiningIn Socio-Academic Perspective‖ International Journal Of Scientific & Technology Research Volume 2, Issue9,
September 2013 Issn 2277-8616.