Optimized PCA Transformation Leveraging Efficiency of Clustering Techniques

Syed.Karimunnisa[1],

Department of Computer Science and Engineering, Koneru Lakshmaiah

Education Foundation Vaddesvaram, Guntur, AP, India-522302, karimun1.syed@gmail.com

supriyamenon05@gmail.com

Supriya Menon M[2]

Department of Computer Science and Engineering, Koneru Lakshmaiah

Education Foundation Vaddesvaram, Guntur, AP, India-522302, karimun1.syed@gmail.com

supriyamenon05@gmail.com

Abstract:

Data mining is the continuous process of finding patterns in large datasets kept in data warehouses. These datasets undergo a variety of pre-processing processes in order to improve the findings. Finding important insights in massive datasets is made possible by efficient data transformation. Principal Component Analysis (PCA) and Wavelet Transformations are two of the most influential transformation techniques. In order to refine PCA results by considerably decreasing the dataset's characteristics, this study explicitly investigates the application of PCA on clustered datasets. This feature reduction results in better performance and better results. Furthermore, the report highlights the effectiveness of this method while also outlining possible directions for future development. To speed up decision-making processes, it suggests combining PCA with other data mining approaches like categorization algorithms. To summarise, the study explores how to improve PCA by reducing dataset features in clustered datasets in order to maximise performance. It also raises the possibility of using PCA in conjunction with other data mining techniques to make decisions faster and more effectively.

Keywords: Data mining, Data transformation, Clustering techniques.

1. Introduction

Data mining, a discipline occupying larger space in research area is considered or focused in this paper. The topic of this study is the broad field of data mining, which is a major area of scientific interest. It mainly concentrates on information extraction from large and varied datasets drawn from several fields, assisting in the development of well-informed judgements to successfully solve modern challenges. Data mining starts with the characterization of data items and the discovery of correlations and links between them. Association norms evolve as a result of these operations. These criteria then direct the data items' classification into supervised and labelled classes, or on the other hand, they result in the opposing process of clustering. In the end, this leads to the data being categorised and prepared for visualisation. In turn, the visualised results greatly aid in the decision-making process, enabling more intelligent and sensible choices.

Although the previously mentioned data mining process appears promising, it has some disadvantages that may hinder and complicate it. These challenges include things that can reduce the effectiveness of data mining, such as noise, redundancy, and high dimensionality. These issues require data pre-processing in order to be addressed because they significantly affect the effectiveness of key information extraction from big datasets. Raw data that is directly derived from multiple sources is often riddled with inaccuracies, missing data, and inconsistent information. Additionally, it is usually not well-prepared for effective data mining methods.

Preparing and refining data for mining purposes is a crucial and required operation that is known as data pre-processing. Its main goal is to transform data into an analysis-ready format. This entails steps including shrinking the size of the dataset, exposing data linkages, normalising values, eliminating outliers, extracting important characteristics, and even creating new ones. Pre-processing techniques include a range of approaches, including reduction, integration, cleansing, and transformation of data. Other preprocessing processes, especially transformation combined with reduction, provide a wider research field than cleaning and integration. This work conducts a survey of the current methods, concentrating on PCA transformation as opposed to wave transforms.

2. Related Work

Tanveer Jahan, G. Narasimha, and C.V. GururRao (2012) went into great length about how to apply clustering for privacy-preserving data mining on distorted data. They presented an enhanced version of the SVD system, called Sparsified Singular Value Decomposition (SSVD). Their method, which proved to be quite helpful for protecting datasets of a medical nature, entailed encrypting data using special characters and ASCII codes. A. Srivastava and G. Srivastav (2015) introduced an alternative method that focuses on protecting private medical data in privacy-preserving data mining via transformations. Their approach produced remarkably accurate findings by protecting individual private data in E-Health records using K-Anonymity approaches.

A geometric data perturbation methodology and a random response method were presented by Yifeng Xu and Jie Liu (2010). Compared to other strategies that were available at the time, this method offers improved privacy protection and is specifically designed to handle continuous-type datasets.

A approach combining a random response methodology and geometric data perturbation was presented in 2010 by Jie Liu and Yifeng Xu. Their research is not appropriate for categorical data because it only focuses on numerical datasets. Notwithstanding this drawback, the method has produced impressive accuracy results, surpassing the state-of-the-art data perturbation techniques at the time. Furthermore, H. Chhinkaniwala and S. Garg discussed a variety of methods and difficulties related to privacy-preserving data mining in 2011. In addition to critically analysing the shortcomings of the current approaches for privacy preservation in data mining, their work concentrated on creating a taxonomy of various privacy-preserving data mining methodologies.

In order to evaluate privacy-preserving data mining, M. Reza and Somayyeh Seifi (2011) introduced novel clustering approaches that applied data perturbation techniques. Their research delves into two approaches, namely Naïve Based and choice models, emphasising the unique methodologies employed in each. The writers were able to consistently strike a compromise between protecting the data's privacy and maintaining its usefulness.

H. Chhinkaniwala and S. Garg presented an effective Multiplicative perturbation method centred on tuple value in 2013. They assessed recall and precision metrics and used the K-Means Clustering technique to improve accuracy.

Furthermore, in the field of privacy-preserving data mining, Mr. Kiran Patel presented a novel method for categorizing data streams that same year. Data stream mining and preparing the data were the two main processes in this approach. The Hoeffding technique, designed to minimize information loss, was introduced by the approach.

Data mining algorithms perform better and are more accurate thanks to a data transformation methodology that was proposed by G. Manikandan et al. in 2013. This method makes use of normalization techniques. Furthermore, Tarique Ahmad et al. described a min-max normalization-based strategy in 2014 to safeguard the privacy of sensitive characteristics within a dataset. Prior to starting the data mining process, this approach applies min-max normalisation to the original dataset values. The suggested k-means algorithm successfully maintained both accuracy and privacy, according to experimental data.

A clustering approach was presented by Patel Brijal et al. in 2015 with the goal of protecting sensitive attributes in a dataset. Their approach produced excellent data mining results while reducing information loss by effectively guaranteeing the preservation of important information inside the dataset. Moreover, Anjana Patel et al. demonstrated the use of the k-means clustering algorithm to apply geometric data perturbation to modified and randomized data in 2016. Their suggested approach was made with accuracy and correctness validation in mind, with the goal of obtaining satisfactory accuracy outcomes. The results of the experiments corroborated the claim that their suggested approach could recover the original data values with the least amount of unnecessary information lost.

3. Proposed Methodology

There is a noticeable tendency in mining scenarios to take into account a small number of factors out of the many that are available. By drastically shrinking the feature space, this method seeks to reduce the number of links between variables. This reduction, also known as "dimensionality reduction," aids in preventing model overfitting. While several methods for dimensionality reduction have been developed, they are generally classified as either feature extraction or feature deletion.

All variables are eliminated except for those that have the strongest predictive power for the intended results. This eliminates some of the benefits provided by the variables that were deleted, but it also makes the selected variables easier to interpret. Conversely, feature extraction mitigates this risk by generating new, independent variables while maintaining the core of the original variables. With deliberate design, these new variables are intended to forecast dependent variables. On the other hand, it is unclear where the reduction actually appears. By eliminating the least significant variables while keeping the core elements of our data, the reduction is accomplished. The Principal Component Analysis (PCA) method is one such feature extraction methodology.

In contrast to Linear Discriminant Analysis (LDA), PCA is a dimensionality reduction algorithm. LDA creates new features by using class information, whereas PCA concentrates on variance. As a result, supervised LDA becomes an attractive substitute for unsupervised PCA. By prioritizing an orthogonal transformation, this statistical transformation method transforms correlated variables into linearly uncorrelated ones. It seeks to determine a schedule of linear combinations that maximize the variance of mutually uncorrelated variables and offers a low-dimensional projection.

There are instances where the perception regarding PCA is misleading, as it's commonly believed that PCA selects a few features and discards the rest. In reality, PCA reconstructs a new feature set from the existing data, projecting it onto a new space through a linear transformation. Research studies have demonstrated a significant enhancement in the data space. For instance, in cases where 18 features are initially considered, PCA can effectively reduce them to an 8-feature set with just a minimal 2% information loss.

PCA is believed to improve results, especially when used in conjunction with clustering techniques, such as K-means, which reduces noise. When this transformation methodology is combined with other mining methodologies, such classification, the goal of lowering characteristics and obtaining better results becomes achievable. Combining PCA transforms with classification produces better results. Below is a brief summary of a few categorization methods that work well in conjunction with PCA to provide improved data analysis results.

PCA ensures better results when applied to clustered data, especially in unsupervised classes. For instance, PCA reduces features, gets rid of redundancy, and presents the data with a smaller set of features when used on sample datasets about vehicles, which include 18 attributes like size, colour, circularity, compactness, number of seats, number of doors, trunk size, etc. should be clear and concise. Show only the most significant or main findings of the research. Discussion must explore the significance of the results of the work.

Enhanced PCA Algorithm:

Step 1: Consider the n-dimensional cluster C1.

Step 2: For each feature in cluster C1, find the mean vector.

Step 3: Create Cluster C1's covariance matrix.

Step 4: Determine the corresponding Eigenvalues (v1, v2, v3, … vn) for the Eigenvectors (e1, e2, e3, … en).

Step 5: Using their Eigenvalues as a guide, arrange the Eigenvectors in decreasing order.

Step 6: To generate a n x k dimensional matrix Z, select the k Eigenvectors with the highest Eigenvalues.

Step 7: To project the sample into a new subspace, use Z.

4. Results and Discussions

The results of the Matlab-performed Enhanced PCA are displayed in the graph. The performance is shown on the y-axis, and the features are represented on the x-axis. The graph, which shows enhanced performance by lowering characteristics inside these clusters, is based on three datasets (clusters).

Table 1 software requirements

Parameter	Principal components
Environment	Matlab
Datasets	3
Subplots	2,2,2
Features	18---8
Simulation time	30 sec
Computer	Intel i3, 16GB Ram
Platform	Windows 7, mat lab

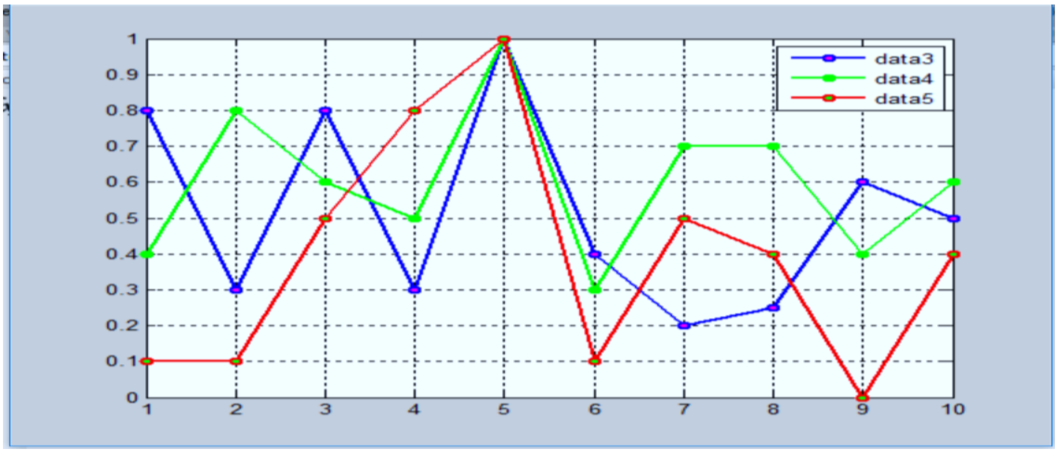

Figure 1 Improved performance Graph

5. Conclusion

Enhanced performance is achieved by combining data transformation with embedded clustering. In the context of data mining, this work focuses on data transformation, with a specific emphasis on clustered data. The importance of safe data transformation for end users without sacrificing information integrity is emphasized throughout the article. The research presents the PCA technique in conjunction with a few clustered datasets in data mining and shows how to project the data in a way that increases mining efficiency. The offered graphs demonstrate the differences in data transformation, suggesting that the suggested techniques could be used to a range of dataset samples. The study recommends using PCA in addition to well-known data mining methods, particularly when it comes to using probability-based algorithms for classification like the probability-based Naïve Bayes algorithm. This integration has the potential to open up new avenues for mining, especially in the medical field, which is expected to be a lucrative sector for future research.

References

1. Km. Swati, Dr. Sanjay Kumar " A Comparative Study of Various Data Transformation Techniques in Data Mining International Journal of Scientific Engineering and Technology " (ISSN: 2277-1581) Volume No.4 Issue No.3, pp : 146-148

2. C. Gokulnath, M.K. Priyan, E. Vishnu Balan, "Preservation of Privacy in Data Mining by using PCA Based Perturbation Technique" 2015 International Conference on Smart Technologies and Management India. 6 - 8 May 2015. Pp: 202-206.
3. Saritha K Sajimon Abraham "Big data challenges and issues: review on analytic techniques" Indian Journal of Computer Science and Engineering (IJCSE)ISSN: 0976-5166 Vol. 8 No. 3 Jun-Jul 2017
4. Ajmeera Kiran Data Mining: "Random Swapping based Data Perturbation Technique for Privacy Preserving in Data Mining" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June: 2019
5. Anastasiya Doroshenko "Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset" IEEE Second International Conference on Data Stream Mining & Processing August 21-25, 2018,
6. Putri, Awalia W Laksmiwati Hira Hybrid "Transformation in Privacy-Preserving Data Mining" 978-1-5090-5671-2/16/©2016 IEEE
7. Chanchal Yadav, 2Shuliang Wang , 3Manoj Kumar "Algorithm and approaches to handle large Data-A Survey" IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013 ISSN (Online) : 2277-5420
8. S.VijayaraniDr.A.Tamilarasi "Data Transformation Technique for Protecting Private Information in Privacy Preserving Data Mining Advanced Computing" : An International Journal (ACIJ), Vol.1, No.1, November 2010
9. N. Sivaram K. Ramar "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5, July 2010
10. Nilesh Kumar Dokania "comparative study of various techniques in data mining international journal of engineering sciences & research technology" ISSN: 2277-9655 7(5): May, 2018
11. Yiyu Yao & Yan Zhao "Explanation-Oriented Data Mining Encyclopedia of Data Warehousing and Mining", 1st edition, Wang, J. (Ed.), 492-297, Idea Group Inc., 2005
12. Uma K, M. Hanumanthappa "Data Collection Methods and Data Pre-processing Techniques for Healthcare Data Using Data Mining" International Journal of Scientific & Engineering Research Volume 8, Issue 6, June-2017 1131 ISSN 2229-5518

13. R.Tamilselvi1, .SivasakthiR.kavitha an efficient preprocessing and post processing Techniques in data mining International journal of research in computer applications and robotics Vol.3 Issue.4, Pg.: 80-85 April 2015
14. Manikandan S "Data transformation Journal of Pharmacology & Pharma cotherapeutics"| July-December 2010 | Vol 1 | Issue 2
15. ShichaozhangAndChengqi Zhang "Data Preparation For Data Mining Applied Artificial Intelligence", 17:375–381, 2003