A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning

Article *in* International Journal of Innovative Research in Science Engineering and Technology · June 2023

DOI: 10.5281/zenodo.8002436

Jasmin Bharadiya
jasminbharadiya

ISSN No:-2456-2165

A Tutorial on Principal Component Analysis for Dimensionality Reduction in Machine Learning

Jasmin Praful Bharadiya
Doctor of Philosophy Information Technology
University of the Cumberlands, USA

Abstract:- Anomaly detection has become a crucial technology in several application fields, mostly for network security. The classification challenge of anomaly detection using machine learning techniques on network data has been described here. Using the KDD99 dataset for network IDS, dimensionality reduction and classification techniques are investigated and assessed. For the application on network data, Principal Component Analysis for dimensionality reduction and Support Vector Machine for classification have been taken into consideration, and the results are examined.. The result shows the decrease in execution time for the classification as we reduce the dimension of the input data and also the precision and recall parameter values of the classification algorithm shows that the SVM with PCA method is more accurate as the number of misclassification decreases. Enormous data in health research is extremely interesting since data-based studies may move more quickly than hypothesis-based research, despite the fact that enormous databases are becoming common and hence challenging to interpret. Using Principal Component Analysis (PCA), one may make some datasets less dimensional. enhances interpretability while retaining most of the information. It does this by introducing fresh variables that are unrelated to one another.

Keywords:- *Machine Learning, Principal Component Analysis, Dimensionality Reduction, Intrusion Detection, Anomaly Detection, Principal Component Analysis, Support Vector Machine.*

I. INTRODUCTION

Machine Learning (ML) is the automatic training of a computer for specific tasks through algorithms. Automatically learning to know user preferences, application sets of algorithms are used to mine data that discover and filter general rules in large data sets. Big data is an environment in which approaches are extract, information is regularly collected or otherwise stored in data sets that are too much or complicated to be managed by standard data processing application. Current use of the term big data uses predictive analytical systems, user behavioral analytics or some other advanced methods of data analysis that extract value from big data and rarely a particular data set size. In many ways, big data deposits have been constructed by corporations with special needs data processing like operating systems has gradually expanded in sizes and number of available data sets, will need massively parallel software running on tens, hundreds, or even thousands of servers to run parallel software.

Through the provision of personalised healthcare and prescriptive analysis, clinical risk response and predictive analysis, reduction of duplication and care variability, automatic external and internal patient reporting, structured medical terminology, and patient registers, big data technology has continued to improve medical treatment. According to particular specified requirements, dimension reduction algorithms convert data from higher dimensions to lower dimensions. Principal Component Analysis (PCA) is a dimensionality-reduction (DR) approach that is primarily used to condense a huge set of variables into a manageable number while retaining the majority of their information.

As computer networks become more prevalent in today's society, so are the dangers that they face. Intrusion detection systems are required to identify numerous threats. Both host-based and network-based intrusion detection systems are available. Host-based technology looks at things like which files were accessed and which programmes were run. With the use of network-based technologies, events are examined as computer-to-computer information packet transfers. Building useful behaviour models to discriminate between normal and pathological activities by watching data is one of the primary challenges for NIDSs.

There are two types of intrusion detection approaches, misuse detection where we model attack behaviour or features using intrusion audit data and anomaly detection, which is to model normal usage behaviours. Usually in the commercial NIDS, the signature or misuse based approach is followed but anomaly based approach is efficient using the machine learning methods. There are many data mining and machine learning methods used for network intrusion detection. Unsupervised methods such as clustering and supervised methods such as Naïve Bayes, Support Vector Machine are used. But comparisons of the results of using an unsupervised dimensionality reduction method along with the supervised

SVM method to SVM without dimensionality reduction is not considered much.

II. SYSTEM DESIGN

The first intrusion detection model based on data mining was proposed by Denning [1] and many research works have been devoted to the construction of effective intrusion detection models. The KD99 dataset for intrusion detection is meant for data mining algorithms, and was established by the Third International Knowledge Discovery and Data Mining Tools Competition [4]. In the KDD99 data set, each data record corresponds to a set of derived features of a connection in the network data. Each connection is labelled either as normal or as an attack, with exactly one specific attack type. In this paper, we will compare various machine learning algorithms that can be used for anomaly detection. The technical challenges in NIDSs based on machine learning methods are dimensionality reduction and classification. There are three main parts depicted in a framework in figure 1 for an intrusion or anomaly detection tool: pre-processing of network data, feature extraction, classification. After the dimension reduction using PCA, the reduced set of features that are linear combination of original features is obtained. The classifier solves the anomaly detection problem using the Support Vector Machine algorithm with the output from the PCA algorithm. Comparison of the classification output using support vector machine (SVM) without dimension reduction and classification using the original data is also done.

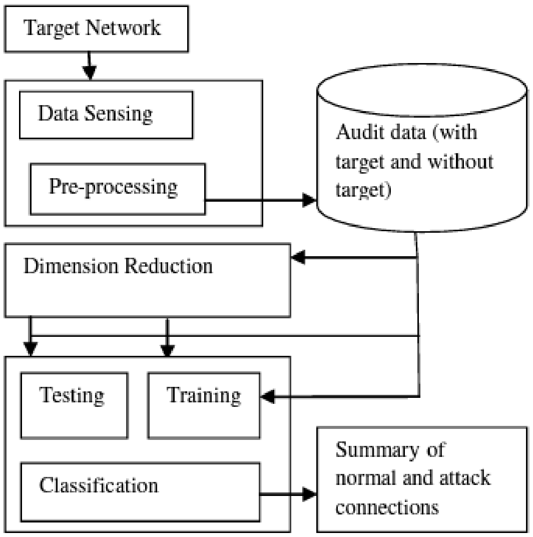

Fig 1 System Design

➢ *Algorithm*

- Consider the network data corresponding to each connection record after mapping. Thus each column represents a dimension of the input data.
- Compute the mean for each dimension, and subtract it from each data value. Compute the covariance matrix C of the input data matrix.
- Calculate the Eigen values and the corresponding eigenvectors for this covariance matrix, and the principal components are computed by solving the eigenvalues problem of covariance matrix
- To find the principal components, choose the eigenvectors corresponding to K largest eigenvalues, where K<<N. Dimensionality reduction step keep only the terms corresponding to the K largest eigenvalues. Hence obtain a new feature vector consisting of eigenvectors of principal components. The final data computed using this feature vector and the mean adjusted original input data using the given equation
- $Final\ Data = RowFeatureVector \times RowDataAdjust$
- Row Feature Vector is the matrix in which eigenvectors in the columns transposed and Row Data Adjust is the mean adjusted input data. The obtained subspace is spanned by the orthogonal set of Eigen vectors, which reveal the maximum variance in the data space.

➢ *Dimensionality Reduction Techniques (DR)*

Dimensionality Reduction (DR) algorithm, which aims to reduce the distance in a latent space between distributions of different data sets to allow efficient transfer learning. The results point that for each device individually, the findings with Dimensionality Reduction (DR) are much preferable to those without decreased dimensionality.The low dimensional data representation of the initial data tends to overcome the issue of the dimensionality curse, and can be easily analyzed, processed, and visualized. Advantage of dimensionality reduction techniques applied to a dataset.

- Decrease the number of dimensions, and data storage space.
- It requires less time to compute.
- Irrelevant, noisy, and redundant data can be deleted.
- Data quality may well be optimized.
- Helps an algorithm to work efficiently and improves accuracy.
- Allow to visualize data
- It simplifies the classification and increases performance as well.

Generally, the DR techniques are classified into two main different techniques: Feature Selection (FS) and Feature Extraction (FE). Feature Selection (FS) is considered an important method since data is constantly produced at an ever-increasing rate; with this method, some significant dimensionality concerns can be minimized, such as effectively decreasing redundancy, eliminating unnecessary data, and enhancing comprehensibility of results. Moreover, Feature Extraction (FE) addresses the issue of finding the most distinctive, informative, and reduced set of features to improve the efficiency of both data processing and storage.

- *PCA in a streaming setting*

A streaming setting in PCA is characterized by sequentially arriving data points over a period of time during which the parameters describing the subspace are repeatedly updated. Over a period of time, the covariance matrix or the subspace can vary, so that tracking and reacting to such changes is necessary to maintain a best possible approximation. PCA algorithms capable of updating its set of parameters continuously without knowledge of the history of data are referred to as online PCA. Popular types of algorithms that fall under the term of online PCA are: incremental PCA or incremental SVD and neural network-based PCA, the focus of this work lies on the latter.

- *Neural network-based PCA.*

Neural network-based PCA refers to typically unsupervised methods that estimate the eigenvalues λi and eigenvectors wi online from the input data stream . These methods are particularly useful for high-dimensional data streams since they avoid the computation of the large covariance matrix. In addition, they can track non-stationary data (i.e. data with a slowly changing covariance matrix). While the development of neural network-based PCA is described in the previous section, it is the focus of this section to provide a more technical view of the neural network-based PCA that is extended and benchmarked in this work. The PCA extended in this work by an adaptive dimensionality adjustment is based on a robust recursive least square algorithm (RRLSA) with interlocking of learning and Gram-Schmidt orthonormalization. In this method, the eigenvectors are updated in a hierarchically way: The eigenvector with the largest eigenvalue is obtained using a single-unit learning rule applied to the original data.

III. APPLICATIONS OF PCA

- **Image and Video Processing:** PCA is widely used in image and video processing tasks such as face recognition, image compression, and denoising. By reducing the dimensionality of image data, PCA can effectively capture the most important features and patterns.

- **Signal Processing:** In signal processing, PCA can be used for feature extraction, noise reduction, and signal classification. It helps in identifying the underlying structure and relevant features in signals

- **Genomics and Bioinformatics**: PCA finds applications in genomics and bioinformatics for analyzing gene expression data, identifying disease subtypes, and understanding the relationships between genes. It aids in identifying significant features and reducing noise in high-dimensional biological datasets.

- **Financial Data Analysis**: PCA is applied in financial data analysis to identify patterns and reduce the dimensionality of financial time series. It helps in portfolio optimization, risk assessment, and identifying influential factors in financial markets.

- **Text Mining:** PCA can be used in text mining to analyze large text datasets, such as document collections or social media data. It aids in feature extraction, topic modeling, and sentiment analysis.

IV. ADVANCEMENTS IN PCA:

- **Kernel PCA**: Kernel PCA extends PCA to nonlinear dimensionality reduction by employing a kernel function to map the data into a higher-dimensional feature space. It allows capturing complex relationships between variables and is particularly useful when the data has nonlinear structures.

- **Sparse PCA**: Sparse PCA incorporates sparsity constraints into the PCA framework, promoting the identification of a sparse set of principal components. This is beneficial when dealing with high-dimensional data where only a few variables contribute significantly to the data structure.

- **Incremental PCA:** Incremental PCA enables the application of PCA on large datasets that cannot fit into memory. It processes data in batches or incrementally updates the principal components, making it more computationally efficient and scalable.

- **Robust PCA**: Robust PCA is designed to handle outliers and noise in the data. It separates the data into low-rank and sparse components, effectively extracting the underlying structure even in the presence of outliers or corrupted data

- **Online PCA**: Online PCA adapts PCA for streaming data where new observations arrive continuously. It updates the principal components incrementally, allowing real-time analysis and adaptive dimensionality reduction.

These advancements in PCA techniques have expanded its applicability and improved its performance in various domains, addressing specific challenges and requirements of different datasets and scenarios.

V. BENEFITS OF PRINCIPAL COMPONENT ANALYSIS (PCA):

- **Dimensionality Reduction**: PCA helps in reducing the dimensionality of high-dimensional datasets by identifying a smaller set of principal components that capture the most important information in the data. This reduces computational complexity, memory requirements, and improves algorithm efficiency.

- **Feature Extraction**: PCA can extract meaningful features from complex datasets, allowing for better understanding and interpretation of the underlying data structure. It helps in identifying the most influential variables or features contributing to the variation in the data.

- **Noise Reduction:** PCA can effectively filter out noise and irrelevant variations in the data by focusing on the components with the highest eigenvalues. It helps in improving the signal-to-noise ratio and enhances the performance of subsequent analysis or modeling tasks.

- **Visualization**: PCA enables the visualization of high-dimensional data in a lower-dimensional space. By projecting the data onto a reduced set of principal components, it allows for the visualization of clusters, patterns, and relationships in the data, aiding in exploratory data analysis.

- **Multicollinearity Detection:** PCA can identify and address multicollinearity issues in datasets, where variables are highly correlated. It helps in identifying linear dependencies among variables and provides a more independent set of components.

VI. LIMITATIONS OF PRINCIPAL COMPONENT ANALYSIS (PCA):

- **Linearity Assumption:** PCA assumes that the data is linearly related to the principal components. If the underlying data has complex nonlinear relationships, PCA may not capture all the relevant information, and other nonlinear dimensionality reduction methods may be more appropriate.

- **Loss of Interpretability:** While PCA reduces the dimensionality of the data, the resulting principal components are usually combinations of the original variables, making their interpretation less straightforward. The interpretability of the transformed features may be challenging, especially when dealing with a large number of components.

- **Sensitivity to Outliers:** PCA is sensitive to outliers in the data, as outliers can disproportionately influence the estimation of principal components. Outliers can distort the resulting variance-covariance structure and affect the quality of dimensionality reduction.

- **Information Loss:** PCA aims to capture the most important information in the data, but there is inevitably some loss of information during the dimensionality reduction process. The lower-dimensional representation may not fully retain all the details and nuances present in the original data.

- **Selecting the Number of Components**: Determining the optimal number of principal components to retain is a subjective decision. Choosing too few components may result in significant information loss, while retaining too many components may lead to overfitting or unnecessary complexity in the data representation.

It is important to consider these limitations and assess the suitability of PCA for specific datasets and analysis goals. In some cases, alternative dimensionality reduction techniques or customized approaches may be more appropriate to address specific challenges or requirements.

- **Considerations and Future Directions:**

The article emphasizes important considerations when applying PCA, such as selecting the appropriate number of principal components, assessing the quality of dimensionality reduction, and evaluating the impact on downstream tasks. It also identifies open research challenges, such as incorporating domain knowledge into PCA, handling high-dimensional and streaming data, and addressing the interpretability of the reduced feature space.

VII. CONCLUSION

In conclusion, Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction in machine learning and data analysis. It offers several benefits, including dimensionality reduction, feature extraction, noise reduction, visualization, and multicollinearity detection. PCA can help in simplifying complex datasets, improving computational efficiency, and enhancing the interpretability of data.

However, PCA also has certain limitations that need to be considered. It assumes linearity, which may not hold in all cases, and its interpretability can be challenging due to the combination of original variables in the transformed components. PCA is sensitive to outliers and may lead to information loss during the dimensionality reduction process. Additionally, determining the optimal number of components to retain requires subjective decision-making.

Overall, PCA is a valuable tool for dimensionality reduction, particularly in cases where linearity is a reasonable assumption and interpretability is not the primary concern. It is important to understand the benefits and limitations of PCA and consider alternative methods when necessary to address specific challenges or requirements in data analysis.

REFERENCES

[1]. Aoying Zhou, Zhiyuan Cai, Li Wei, Weining Qian. M-kernel merging: towards density estimation over data streams. In: Eighth International Conference on Database Systems for Advanced Applications, 2003. (DASFAA 2003). Proceedings.; 2003. p. 285–292.

[2]. https://ijisrt.com/convolutional-neural-networks-for-image-classification

[3]. Artac M, Jogan M, Leonardis A, “Incremental PCA for on-line visual learning and recognition,” Object recognition supported by user interaction for service robots. 2002;3:781-784.

[4]. Lee JA, Verleysen M. Nonlinear Dimensionality Reduction. Springer-Verlag GmbH; 2007.

[5]. W.K. Lee, et al., "Mining audit data to build intrusion detection models", In Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pp.66-72, 1998.

[6]. Zhangxue-qin, Gu chun-hua and Linjia-jun, “Intrusion detection system based on feature selection and support vector machine”, East China University of Science and Technology, Proceedings of IEEE, 2006.

[7]. M. M. Sebring, E. Shellhouse, M. E. Hanna, and R. Alan Whitehurst, “Expert systems in intrusion detection: A case study”, In Proceedings of the 11th National Comput Security Conference, Baltimore, Maryland.

[8]. Jasmin Praful Bharadiya https://journaljerr.com/index.php/JERR/article/view/858

[9]. CHEN Bo, Ma Wu, “Research of Intrusion Detection based on Principal Components Analysis”, Information Engineering Institute, Dalian University, China, Second International Conference on Information and Computing Science, 2009.

[10]. Jasmin Praful Bharadiya https://journalajorr.com/index.php/AJORR/article/view/164

[11]. Bannour S, Azimi-Sadjadi MR. Principal component extraction using recursive least squares learning. IEEE Transactions on Neural Networks. 1995;6(2):457–469. pmid:18263327

[12]. Zhang T, Yang B. Big Data Dimension Reduction Using PCA. In: 2016 IEEE International Conference on Smart Cloud (SmartCloud); 2016. p. 152–157.

[13]. Oja E. Simplified neuron model as a principal component analyzer. Journal of Mathematical Biology. 1982;15(3):267–273. pmid:7153672