Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Investigating the Influence of Random Projection on Classification

[1]Dr. Raghunadh Pasunuri, [2]Dr. Rajaram Jatothu

[1]PDF Scholar, Computer Science & Engineering, Institute of Engineering & Technology, Srinivas University, Mangalore, Karnataka, India.

[2]Professor, Dept. of CSE, Malla Reddy Engineering College for Women, Hyderabad, India.

ARTICLEINFO

Received: 01 Nov 2024

Revised: 14 Dec 2024

Accepted: 26 Dec 2024

ABSTRACT

Dealing with High-dimensional data is very common in this digital era. The high-dimensionality poses certain challenges like, we cannot apply the algorithms and the accuracy of clustering or classification algorithms will be reduced because of the high dimensionality of the data. Many works have been proposed to reduce the dimensionality and apply the classification and clustering algorithms on the reduced data, without losing the essential information from the data. In this work, we want to apply projection-based dimensionality reduction (DR) techniques as a pre-processing step, before applying the classification or clustering algorithm on the data.

Keywords- High-dimensional data, Dimensionality reduction, Random projection, Classification, Feature extraction

INTRODUCTION:

High-dimensional data has become ubiquitous in modern applications driven by rapid advancements in digital technologies, including domains such as bioinformatics, image processing, text analytics, and social networks. While the availability of large numbers of features can potentially provide richer information, it also introduces significant challenges for data analysis. One of the fundamental issues associated with high-dimensional data is the well-known curse of dimensionality, where the volume of the feature space increases exponentially with dimensionality. This leads to data sparsity, increased computational complexity, and degraded performance of conventional machine learning algorithms.

In particular, classification and clustering algorithms often struggle to perform effectively in high-dimensional spaces, as the distance measures and similarity metrics they rely upon become less meaningful. As a result, the accuracy and reliability of these algorithms tend to decline when applied directly to such data. Moreover, the presence of redundant and irrelevant features further exacerbates the problem, making it difficult to extract meaningful patterns and insights.

To address these challenges, dimensionality reduction (DR) techniques have been widely explored as a crucial preprocessing step in data analysis pipelines. These techniques aim to transform high-dimensional data into a lower-dimensional representation while preserving its intrinsic structure and essential information. Among various approaches, projection-based dimensionality reduction methods have gained significant attention due to their effectiveness and computational efficiency. Techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) project data onto a lower-dimensional subspace that captures the most informative aspects of the original dataset.

In this context, the present work focuses on the application of projection-based dimensionality reduction techniques as a preprocessing step prior to performing classification and clustering. By reducing dimensionality while retaining critical information, the proposed approach aims to enhance the performance, accuracy, and efficiency of subsequent learning algorithms. The study seeks to demonstrate that appropriate dimensionality reduction can mitigate the adverse effects of high dimensionality and lead to more robust and interpretable results in data analysis tasks.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LITERATURE REVIEW:

Earlier, many researchers have been presented many reviews on Dimensionality Reduction Techniques (DRT's) for various applications and for different data sets.

Zaheer et al., [1] proposed an approach for time series data analysis by combining the strengths of various dimensionality reduction methods with robust machine learning models. They applied PCA, Kernel PCA, Sparse PCA, Incremental PCA, ICA, SVD, t-SNE and UMAP to reduce the dimensionality of the data set without losing the essential information of the data. After the DRT is applied to the data set and then a machine learning model like ARIMA, KNN, Random Forest, LASSO and XGBoost is applied on the reduced data set. The proposed approach not only improves the predictive accuracy but also provides a scalable solution for handling large and complex datasets. This study shows the importance of dimensionality reduction in managing the complexity of time series data.

S Ravindran and G Aghila [2] proposed a Data Independent Reusable Projection (DIRP) technique that projects the high-dimensional data into low-dimensional data and proved that the projection matrix can be reused for any dataset with same number of dimensions. In this work, the DIRP method was implemented in R and a new package called "RandPro" is created to generate the projection matrix, and it is tested with the CIFAR-10, MNIST handwritten digit recognition data set and human activity recognition (HAR) dataset. A comparative analysis is done on the running time and classification metrics comparison has been done between PCA and DIRP methods. The experimental results shows that the proposed DIRP method significantly reduces the computational complexity with a very small loss of accuracy when compared with all dimensions as well as PCA. The running time of DIRP is also very less when compared to PCA. The main advantage of DIRP method is that there is no need of generating a projection matrix every time, which leads to a data-independent projection matrix technique and makes the DIRP suitable for dimension reduction in big data classification.

S Ravindran and G Aghila [3] used the Data Independent Reusable Projection (DIRP) technique for solving the privacy issue during the data analysis. They used different random distribution methods to generate the projection matrix there by masking the original data with the randomness of the data, which is used to solve the privacy issue in the data analysis task. The proposed algorithm was tested on MNIST handwritten digit recognition dataset, and the experimental results shows that, the proposed algorithm reduces the computational complexity of the feature extraction dimension reduction method and the data privacy is preserved during the analysis phase. The DIRP method reduces the distortion loss in the reduced dimensional space and also the running time.

Recent advances in Machine Learning have led to extensive research on dimensionality reduction (DR) techniques to address the challenges posed by high-dimensional data, particularly the curse of dimensionality. Traditional linear methods such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) continue to serve as foundational approaches due to their simplicity, interpretability, and computational efficiency. However, recent studies have sought to enhance these classical techniques by improving their robustness and interpretability. For instance, Bonetti et al. (2024) proposed an interpretable linear DR framework based on bias–variance analysis, highlighting the importance of balancing model complexity and generalization.

In parallel, supervised and explainable dimensionality reduction methods have gained attention. Björklund et al. (2023) introduced SLISEMAP, a supervised DR approach that integrates local explanations, enabling better interpretability of reduced feature spaces. Such methods are particularly useful in applications where transparency and explainability are critical.

With the increasing complexity of real-world datasets, nonlinear dimensionality reduction techniques have also been widely explored. Methods such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Isomap are effective in capturing intrinsic data structures and manifolds. Recent work by Yousuff et al. (2024) demonstrated the effectiveness of nonlinear DR techniques in handling high-dimensional biological datasets such as single-cell RNA sequencing data. Similarly, Mahapatra and Chandola (2024)

extended manifold learning approaches to non-stationary data streams, emphasizing their adaptability in dynamic environments.

Deep learning-based dimensionality reduction methods have emerged as a powerful alternative to traditional approaches. Autoencoders, particularly sparse and denoising variants, have been widely used to learn compact representations of high-dimensional data. Manjunatha et al. (2024) applied sparse deep denoising autoencoders for intrusion detection, achieving improved performance in high-dimensional cybersecurity datasets. Furthermore, Berahmand et al. (2024) provided a comprehensive survey on autoencoders, highlighting their versatility and effectiveness across multiple domains.

Recent studies have also explored hybrid and neural approaches to dimensionality reduction. Li and Wang (2024) proposed a neural network-based extension of PCA, combining the strengths of linear projection and deep learning. Additionally, multi-stage dimensionality reduction frameworks have been introduced to improve classification performance, as seen in recent work on social media data analysis (2024), where a two-stage DR approach enhanced feature representation and model accuracy.

Comparative and survey-based studies continue to play an important role in evaluating the effectiveness of different DR techniques. Recent comparative analyses (2023) have examined linear, nonlinear, and manifold-based approaches, providing insights into their strengths and limitations across various applications. Emerging research directions also include the application of DR in advanced fields such as quantum machine learning and IoT data analytics, where efficient handling of high-dimensional data remains a critical challenge.

Overall, the literature indicates that while numerous dimensionality reduction techniques have been developed, projection-based methods remain highly relevant due to their scalability, efficiency, and ease of integration with downstream learning algorithms. Building on these advancements, the present work focuses on applying projection-based dimensionality reduction techniques as a preprocessing step to enhance the performance of classification and clustering algorithms in high-dimensional settings.

METHODOLOGY:

Problem Statement:

We want to study the effect of applying the classification algorithm on the data set after applying the random projection dimensionality reduction technique.

Key Objectives of the Study:

1. To test the performance of machine learning algorithms on different data sets before and after applying the dimensionality reduction techniques.
2. Which machine learning algorithm is suits with which type of data, after the dimensionality reduction method application to the original data?
3. How does different dimensionality reduction methods can preserve the essence of the original data after applying the reduction method?

The proposed methodology is shown in the Figure 1:

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

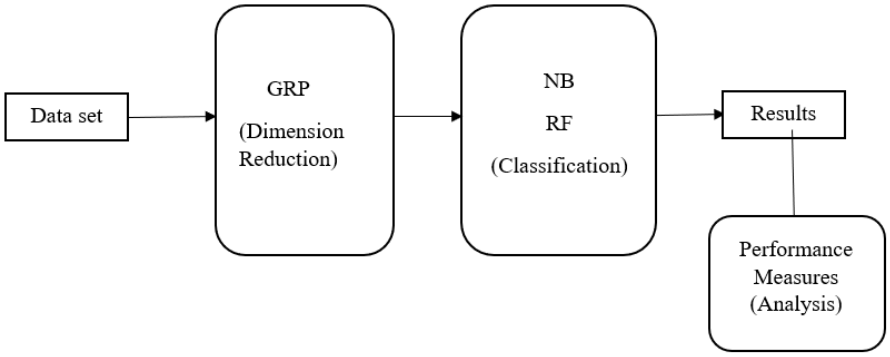

Figure 1: The proposed methodology

ML Algorithms: Clustering and Classification Algorithms

As part of this work, we have taken two classification algorithms into consideration, one is Naïve Bayes and the other one is Random Forest algorithm. The study focuses on studying the effect on dimensionality reduction on the classification performance of the two algorithms.

For reducing the dimensionality of the given data set, we have used the famous data-independent technique called Random Projection. The random projection method takes the input data set as a matrix (Input) and creates

Random Projection matrix generation methods: NRP/GRP, PMRP, HRP, DIRP

Classification Models

This study evaluates four widely used supervised learning algorithms: Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and k-Nearest Neighbors (KNN). These models represent probabilistic, tree-based, ensemble, and instance-based learning paradigms, respectively.

Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes' theorem with a strong conditional independence assumption among features. The posterior probability of a class C given an input feature vector $X=(x_1, x_2, \ldots, x_n)$ is computed as:

$$P(C \vee X) = \frac{P(C)\prod_{i=1}^{n} P(x_i \vee C)}{P(X)}$$

For classification, the decision rule is:

$$\hat{C} = \arg\max_{C} P(C)\prod_{i=1}^{n} P(x_i \vee C)$$

Despite its simplifying assumption of feature independence, Naive Bayes performs efficiently in high-dimensional spaces and is particularly effective for sparse data distributions.

Decision Tree (DT)

A Decision Tree is a non-parametric supervised learning model that recursively partitions the feature space into homogeneous regions. At each node, an optimal split is selected using impurity measures such as Gini impurity or entropy.

The Gini impurity is defined as:

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$G = 1 - \sum_{i=1}^{K} p_i^2$$

Where p_i is the proportion of samples belonging to class i. Decision Trees are highly interpretable but prone to overfitting without pruning constraints.

Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap aggregation (bagging) and random feature selection. The final prediction is obtained via majority voting:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \ldots, T_M(x)\}$$

Where, $T_m(x)$ represents the prediction of the m^{th} tree. Random Forest reduces variance and improves generalization compared to a single Decision Tree.

K-Nearest Neighbors (KNN)

k-Nearest Neighbors is an instance-based learning algorithm that assigns a label to a query point based on the majority class among its k closest training samples. The Euclidean distance is commonly used:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

KNN is non-parametric and requires no explicit training phase, but suffers from high computational cost during inference and sensitivity to feature scaling.

Data sets used: MNIST (60K by 784)

The **MNIST dataset** is a widely used benchmark dataset in machine learning and computer vision for handwritten digit recognition. It consists of gray scale images of digits ranging from **0 to 9**, originally curated from the NIST database and later standardized for research purposes.

In its commonly used training configuration, MNIST contains **60,000 samples**, where each sample represents a handwritten digit image. Each image is normalized to a fixed size of **28 × 28 pixels**, which, when flattened into a feature vector, results in a **784-dimensional input representation** (i.e., (28 by 28 = 784)).

Each of these 784 features corresponds to the pixel intensity value of a gray scale image, typically normalized to the range ([0, 1]) or ([0,255]). The dataset is labeled into **10 classes**, corresponding to the digits 0 through 9, making it a multi-class classification problem.

Formally, the dataset can be represented as:

$$[X \in \mathbb{R}^{60000X784}, Y \in 0,1,..,9^{60000}]$$

Where, (X) denotes the input feature matrix and (Y) denotes the associated class labels.

MNIST is particularly valued due to its:

- Moderate dimensionality (784 features), suitable for testing dimensionality reduction techniques,
- Balanced class distribution across 10 digit categories,
- Clean and preprocessed structure, making it ideal for benchmarking classical and modern classifiers.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Despite its simplicity, MNIST continues to serve as a foundational dataset for evaluating classification algorithms, feature extraction methods, and dimensionality reduction techniques such as Gaussian Random Projection.

PRAGMATIC ANALYSIS AND PERFORMANCE COMPARISON:

Table1: Accuracy and F1-score values for GRP+NB classifier on MNIST

GRP+NB on MNIST 60000 X 784		
Red.Dim(k)	Accuracy (%)	F1-Score (%)
10	57	58
50	77	77
100	81	81
200	83	83
300	83	83
400	84	84
500	84	84
600	84	84
700	84	84
784	84	84

The performance of the classification model using Gaussian Random Projection (GRP) combined with Naive Bayes classifier (NB) was evaluated on the MNIST dataset with varying reduced dimensions. The dataset originally consists of 60,000 samples with 784 features, which were projected into lower-dimensional spaces ranging from 10 to 784.

The results indicate a clear trend in classification performance as the dimensionality increases. At very low dimensions (k = 10), the model achieves an accuracy of 57% and an F1-score of 58%, reflecting significant information loss due to aggressive dimensionality reduction. However, as the number of dimensions increases, there is a substantial improvement in performance. At k = 50, accuracy rises sharply to 77%, demonstrating that even moderate dimensionality can preserve a considerable amount of discriminative information.

Further increases in dimensionality lead to gradual improvements. At k = 100 and k = 200, the accuracy reaches 81% and 83%, respectively. Beyond k = 300, the performance begins to plateau, with accuracy stabilizing around 84%. Notably, increasing the dimensionality beyond 400 up to the original 784 dimensions does not yield significant gains, indicating diminishing returns.

This behavior highlights an important characteristic of projection-based dimensionality reduction: a relatively low-dimensional representation can retain most of the essential structure of the data. The plateau observed beyond k ≈ 300 suggests that additional dimensions contribute minimal new information for classification in this setup.

Another key observation is the consistency between accuracy and F1-score across all dimensions, indicating balanced class predictions and stable model behavior. This is particularly relevant for datasets like MNIST, where class distribution is relatively uniform.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Overall, the results demonstrate that applying Gaussian Random Projection as a preprocessing step can significantly reduce dimensionality (by more than 50%) while maintaining near-optimal classification performance. This not only reduces computational complexity but also mitigates issues related to high-dimensional data, making GRP a practical and efficient approach for large-scale classification tasks.

Table2: Accuracy and F1-score values for GRP+RF classifier on MNIST

GRP+RF on MNIST 60000 X 784		
Red.Dim(k)	Accuracy (%)	F1-Score (%)
10	68	68
50	90	90
100	92	92
200	93	93
300	94	94
400	94	94
500	94	94
600	94	94
700	94	94
784	94	94

Result Analysis: GRP + RF on MNIST (60,000 × 784)

The table reports classification performance of a pipeline combining Gaussian Random Projection (GRP) with a Random Forest (RF) classifier on the MNIST dataset, evaluated across varying reduced dimensionalities k. The metrics considered are Accuracy and F1-score, which are identical here, indicating balanced class performance across digits.

1. Effect of Dimensionality Reduction

A clear trend emerges as the reduced dimension k increases:

- At **very low dimensionality (k = 10)**, performance is relatively weak, with both Accuracy and F1-score at **68%**. This suggests that aggressive compression leads to substantial loss of discriminative information.
- A sharp improvement occurs when increasing to **k = 50**, where performance jumps to **90%**, indicating that most essential structure of the MNIST data is already partially preserved even at moderate dimensions.
- Further increases to **k = 100–200** continue to improve performance gradually, reaching **92–93%**, showing diminishing but still meaningful gains as more feature information is retained.

2. Saturation Behavior

From **k = 300 onward**, performance stabilizes:

- Accuracy and F1-score plateau at approximately **94%**.
- Increasing dimensionality up to the original space (**k = 784**) does not yield any additional improvement.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This indicates that the Random Forest classifier does not benefit from the full high-dimensional representation once sufficient structure has been captured by GRP. The intrinsic dimensionality of MNIST is effectively much lower than 784.

3. Key Insights

- **Optimal trade-off region:** Around **k = 200–300**, the model achieves near-maximum performance while still benefiting from substantial dimensionality reduction (over 60% reduction from original features).
- **Dimensionality redundancy:** The flat performance beyond k = 300 suggests that higher dimensions contain redundant or non-informative features for this classifier.
- **Effectiveness of GRP:** Despite being an unsupervised and random projection method, GRP preserves enough structure for RF to achieve strong performance (~94%), demonstrating robustness of the pipeline.

4. CONCLUSION

The results show that GRP combined with Random Forest is highly effective for MNIST classification even under significant dimensionality reduction. Performance saturates quickly, implying that an intermediate reduced dimension (around 200–300) provides the best balance between computational efficiency and predictive accuracy. Further increasing dimensionality does not improve performance, highlighting redundancy in the original feature space for this task. In future, we want to study the effect of dimensionality reduction on Decision Tree and kNN classifiers.

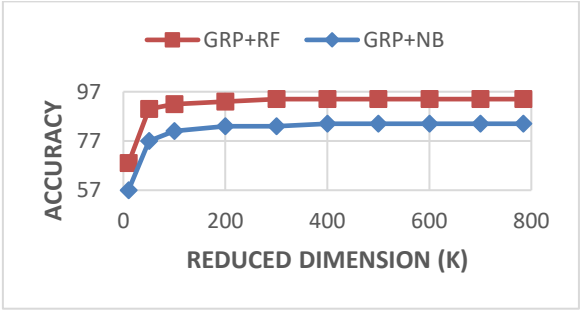

Fig.2: Performance comparison: GRP+NB versus GRP+RF

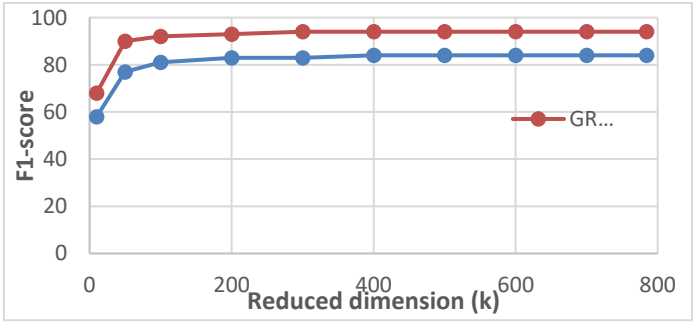

Fig.3: F1-score comparison of GRP+NB versus GRP+RF

CONCLUSION:

This study evaluated the impact of Gaussian Random Projection (GRP) as a dimensionality reduction technique on the MNIST dataset (60,000 × 784) using a Random Forest classifier. The experimental results demonstrate that substantial dimensionality reduction can be achieved without significant loss in classification performance.

The model shows a strong improvement in performance as the reduced dimension increases from 10 to 100, after which the accuracy rapidly stabilizes. Beyond approximately 200–300 dimensions, both accuracy and F1-score plateau at around **94%**, indicating that most of the discriminative information in the dataset is preserved even after significant compression of the original feature space.

These findings highlight that high-dimensional image data such as MNIST contains considerable redundancy, and projection-based dimensionality reduction techniques like GRP can effectively retain essential structural information. Among the evaluated settings, Random Forest proves to be highly robust to dimensionality reduction, maintaining consistent performance even at lower projected dimensions.

Overall, the results confirm that GRP is an efficient and practical preprocessing technique for high-dimensional classification tasks, enabling reduced computational complexity while preserving predictive accuracy.

REFERENCES:

[1] Ravindran Siddharth., and G. Aghila, "A Data-Independent Reusable Projection (DIRP) Technique for Dimension Reduction in Big Data Classification Using k-Nearest Neighbor (k-NN)," Natl. Acad. Sci. Lett. 43, 13–21 (2020). https://doi.org/10.1007/s40009-018-0771-6

[2] Ravindran Siddharth, and G. Aghila. "A Privacy-Preserving Feature Extraction Method for Big Data Analytics Based on Data-Independent Reusable Projection." Handbook of Research on Cloud Computing and Big Data Applications in IoT. IGI Global, 2019. 151-169. http://doi.org/10.4018/978-1-5225-8407-0.ch008

[3] Bonetti, P., Metelli, A. M., & Restelli, M. (2024). Interpretable linear dimensionality reduction based on bias-variance analysis. Data Mining and Knowledge Discovery, 38, 1713–1781. Springer. DOI: 10.1007/s10618-024-01015-0

[4] Björklund, A., Mäkelä, J., & Puolamäki, K. (2023). SLISEMAP: Supervised dimensionality reduction through local explanations. Machine Learning, 112, 1–43. DOI: 10.1007/s10994-022-06261-1

[5] Björklund, A., Mäkelä, J., & Puolamäki, K. (2023). SLISEMAP: Combining supervised dimensionality reduction with local explanations. In Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2022), Lecture Notes in Computer Science (Vol. 13718, pp. 612–616). Springer, Cham. DOI: 10.1007/978-3-031-26422-1_41

[6] Yousuff, M., Babu, R., & Rathinam, A. (2024). Nonlinear dimensionality reduction based visualization of single-cell RNA sequencing data. Journal of Analytical Science and Technology, 15(1), 1–23. https://doi.org/10.1186/s40543-023-00414-0

[7] Mahapatra, S., & Chandola, V. (2024). Learning manifolds from non-stationary streams. Journal of Big Data, 11(1), Article 42. SpringerOpen. DOI: 10.1186/s40537-023-00872-8

[8] Manjunatha, B. A., Shastry, K. A., Naresh, E., Pareek, P. K., & Reddy, K. T. (2024). A network intrusion detection framework on sparse deep denoising auto-encoder for dimensionality reduction. Soft Computing, 28(5), 4503–4517. Springer. https://doi.org/10.1007/s00500-023-09408-x

[9] Berahmand, K., Daneshfar, F., Salehi, E. S., Li, Y., & Xu, Y. (2024). Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review, 57, Article 28. Springer. DOI: 10.1007/s10462-023-10662-6.

[10] Li, J., & Wang, Y. W. (2024). nPCA: A linear dimensionality reduction method using a multilayer perceptron. Frontiers in Genetics, 14, 1290447. https://doi.org/10.3389/fgene.2023.1290447