The Role of Data Normalization in Kernel Methods for Image Classification

Sachin Singh (M24CSE033)

Problem Statement

It can be observed that in the realm of machine learning, particularly in the context of image classification, normalization actually plays a pretty critical role through techniques involving SVMs and Kernel Principal Component Analysis (KPCA). Much as scaling and distribution transformation of input features for efficient utilization of kernel functions is of paramount importance, this resource is generally under-exploited. It is designed to thoroughly and holistically examine how various normalization methods can maximize the performance of such kernel techniques in the context of image classification. Theoretically, developing and testing a learning model for image classification achieves the goal of improving and practically performing applications in the wide range of real-world applications, thereby adequately making a strong contribution in the developments and advancements of machine learning methodologies.

LITERATURE SURVEY

The recent developments in the kernel method of machine learning demonstrate that methodological inventions play important roles in achieving good classification performance, especially for image classification. A seminal survey of the use of positive definite kernels in machine learning is presented by Hofmann, Schölkopf, and Smola, in which they clearly expound upon the wide applicability of the approach from simple binary classifiers to complex structured data analysis.[1] They argue that the flexibility of kernel methods makes them theoretically as well as practically relevant in dealing with nonlinear and non-vectorial data.

Singh and Singh discuss in their work the impact of normalization of data on the performance of classification, which emphasizes preprocessing steps in machine learning.[2] In this research, they discussed fourteen different techniques of normalization, and their effect on feature selection and weighting while performing a classification task. This work is useful in identifying optimum normalization procedures that can improve the accuracy of a machine learning work; however, they have concluded that not one of them dominates all others.

Jian et al proposed a multi-scale learning approach for optimizing the selection of kernel functions in SVMs.[3] Their method incorporates centered polarization in a framework, showing significant improvements toward generalizing over widely different densities of SVM data. Such an approach promises a bright future for enhancements from the kernel methods' view of machine learning.

Camp-Valls and Bruzzone discuss kernel-based approaches to hyperspectral image classification. The authors compare a few of the kernel-based techniques, regularized radial basis function neural networks, and standard SVMs in noisy environments and high-dimensional data spaces to provide hard comparisons.[4] Critical insights are expected to come out from this competition regarding the appropriateness of kernel methods for use in applications based on hyperspectral imaging.

Lastly, Ahmad and Mugdadi finish by including a new approach for normality checking using kernel methods which, in fact, is pretty important to hold the essential assumptions in many statistical models and machine learning algorithms.[5] The suggested procedure which was based on some transformations of the data regarding their independence provided a robust method of statistical testing considering the absence of parameter estimation or transformation, thus making this a good reference for studies that work with data normalization along with the effects it may trigger to the performance of the algorithm.

The work presented here gives an overview of the state-of-the-art techniques in kernel methods and normalization and transformation. The specificity of application, from simple statistical testing to complex image classification, also clearly speaks to the present challenges and opportunities for methodological improvements in machine learning.

Datasets

This dataset of images of skin lesions is comprised of several dermatological studies concentrating on the diagnosis of skin cancers.

Description

There are two types of datasets:

- **Malignant:** Images of skin lesions diagnosed with cancer, to train the models for diagnosis of malignant conditions.

- **Benign:** Pictures of benign skin diseases without cancer, required to train models and differentiate between threatening and harmless conditions.

Challenges

1. **Class Imbalance:** Carcinomas are rarely observed compared to benign lesions, thus resulting in biased model prediction.

2. **Intra-class Variation:** Models become difficult to train with huge intra-class variations.

3. **Inter-class Similarity:** High visual similarity between some benign and malignant lesions challenges accurate classification.

Utility

It is good for binary image classification and suitable for training deep learning models; therefore, it could aid research into cases of early detection and diagnosis of skin cancer with better techniques in image classification.

Objectives

This project focuses on examining how different data-normalization methods can affect the performance of kernel-based machine-learning techniques applied to image classification. In particular, the study intends to:

1. **Increase Classification Accuracy:** The classification accuracy of medical images (benign vs malignant) can be improved through suitable kernel methods and normalization.
2. **Performance of Duration:** Min-Max Scaling and Z-Score Normalization on KPCA and SVM.
3. **Evaluate Model Stability:** Employ cross-validation and tuning parameter(s) for consistent generation of each model for diverse datasets.
4. **Giving Performance Contrast:** Compare diverse normalization methods and kernel algorithms to ascertain the most optimal medical image classification methodology.

Methodology

This project will have four main phases in terms of methodology: **Data Preprocessing**, **Feature Engineering with Kernel Methods**, **Model Selection and Training**, and **Model Evaluation**. The above steps were taken to make sure if the final model was robust and accurate enough to classify between benign images vs malignant.

1. Data Preprocessing

The dataset consists of images of skin lesions, and is classified into two categories:

- **Data Collection**: The dataset comprises images of skin lesions, divided into two categories:
  - **Malignant**: Images labelled with cancer.
  - **Benign**: Images of non-cancerous skin growth.
- **Data Cleaning**: To avoid the lack of quality in data corrupted or irrelevant images are removed from dataset. This approach guaranteed that the images used are consistent for training and testing.
- **Image Resizing**: I resized every image to 128×128 pixels so that all images have the same dimensionality, and this would make it easier for computation.

- **Data Normalization**: The data took two normalization processes to become the form of acceptable data for input:
  - **Min-Max Normalization:** We computed pixel values within range [0, 1]. The model performed better with this technique than the one without this technique because it standardizes the scale of data so that they are closer to each other.
  - **Z-Score Normalization:** Data centered to a mean of 0 with a standard deviation of 1, further improving model final convergence and data stability (especially for Support Vector Machines).

2. Feature Engineering with Kernel Methods

Since image data is high-dimensional and some of the relations are pretty complicated, we rely on Kernel Principal Component Analysis (KPCA):

- **Kernel PCA**: KPCA was utilized for performing dimensionality reduction while using RBF (Radial Basis Function) kernel. We selected RBF kernel since the model must be able to obtain non-linear patterns within the data, which will play an important part in differentiating similar benign and malignant images.
- **Dimensionality Reduction**: KPCA reduced the number of features, while preserving the most discriminative information extracting components supporting the necessary patterns needed for classification. The same also lowers computational costs with high accuracy.

3. Model Selection and Training

The support vector machine (SVM) were selected and trained to test the impact of normalization methods and kernel:

- **Kernel Selection**: An RBF Kernel for SVM was chosen to deal with the non-linear separations of classes.
- **Hyperparameter Tuning**: Hyperparameters like C (regularization parameter) and gamma (kernel coefficient), were tuned using Grid Search, and Cross-validation technique.

4. Model Evaluation

Multiple metrics were used to assess the performance of each model for reliable classification of benign and malignant cases:

- **Cross-Validation**: To test the stability and generalizability, 5-fold cross-validation was used. This technique helped prevent any train-test split from providing biased performance for the model.
- **Evaluation Metrics**:

  - **Accuracy**: The overall correctness of predictions, represented as the ratio of true positive and true negative predictions.

  - **Precision**: The model can predict malignant cases as malignant by reducing false positives.

  - **Recall**: How sensitive the model is for malignant cases, an important metric in medical imaging since it would lead to a missed diagnosis.

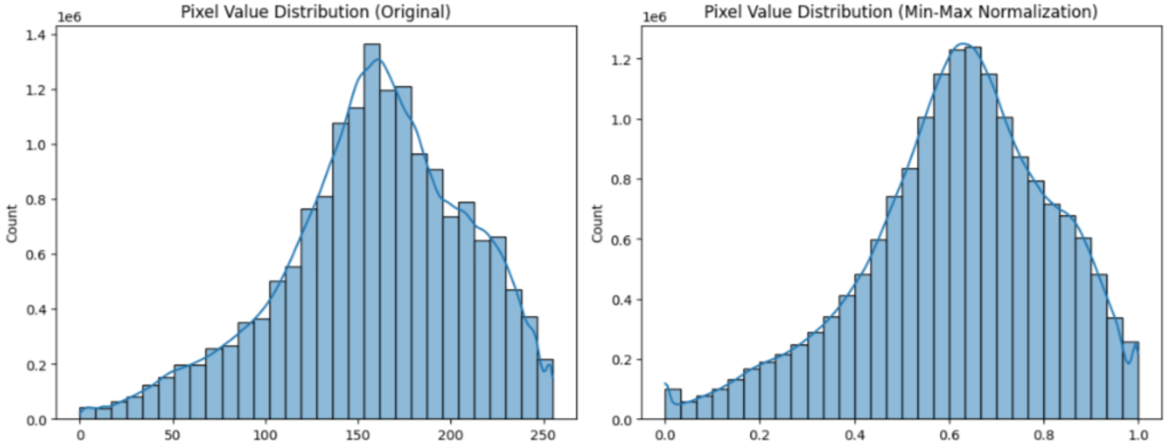
  - **F1-Score**: The harmonic mean of precision and recall tries to find the balance between precision and recall.

- **Confusion Matrix and Visualizations:** A confusion matrix was generated to display the counts of true positive, true negative, false positive, and false negative predictions. Additional bar plots were created for accuracy, precision, recall, and F1-score to allow easy comparison of performance across normalization techniques and model types.

Result

Following the implementation and assessment of the models, the results listed below were achieved.:

1. **Normalization Impact**:



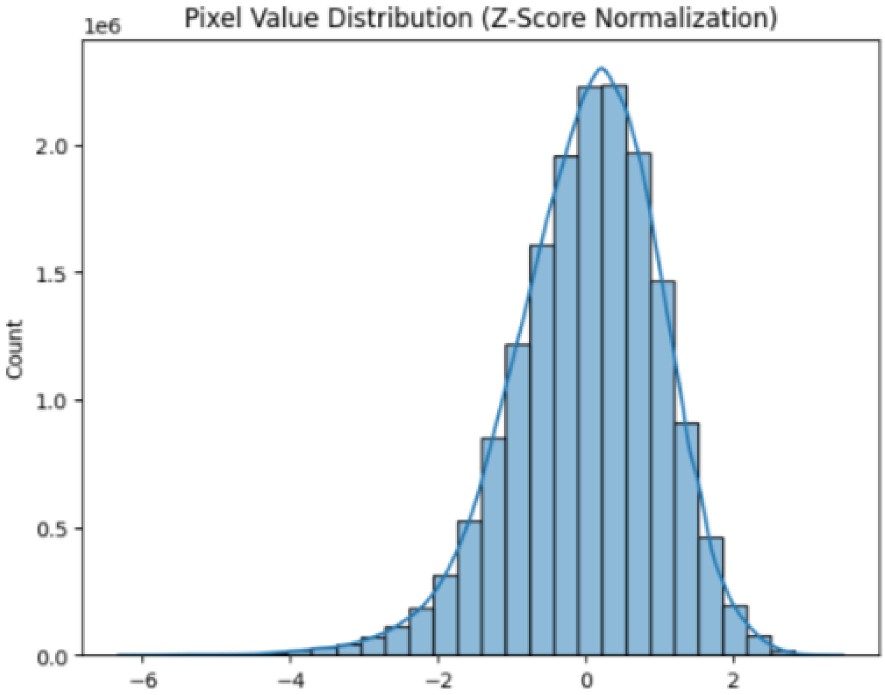


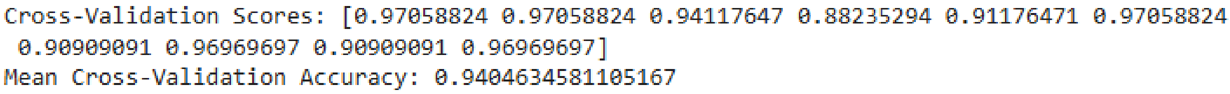
Fig: Pixel Distribution Before and After Min-Max and Z-Score Normalization

- **Min-Max Normalization:** Improved data distribution by scaling values between 0 and 1, allowing the model to process pixel intensities more uniformly.

- **Z-Score Normalization:** Centered the data around a mean of 0 with a standard deviation of 1, further enhancing the SVM model's convergence and performance.

2. **Kernel PCA with RBF Kernel**: Utilizing Kernel PCA with a radial basis function (RBF) kernel allowed us to decrease dimensionality while preserving key characteristics, enabling the model to attain a high recall rate for malignant instances.

3. **Cross-Validation Consistency**: The average accuracy from cross-validation was around 90.6%, showing little variation between folds. This reflects a significant consistency in the performance of the SVM model, indicating robust generalization abilities.



```
Cross-Validation Scores: [0.97058824 0.97058824 0.94117647 0.88235294 0.91176471 0.97058824
 0.90909091 0.96969697 0.90909091 0.96969697]
Mean Cross-Validation Accuracy: 0.9404634581105167
```

Fig: Cross-validation Scores

4. **SVM Model Performance**:

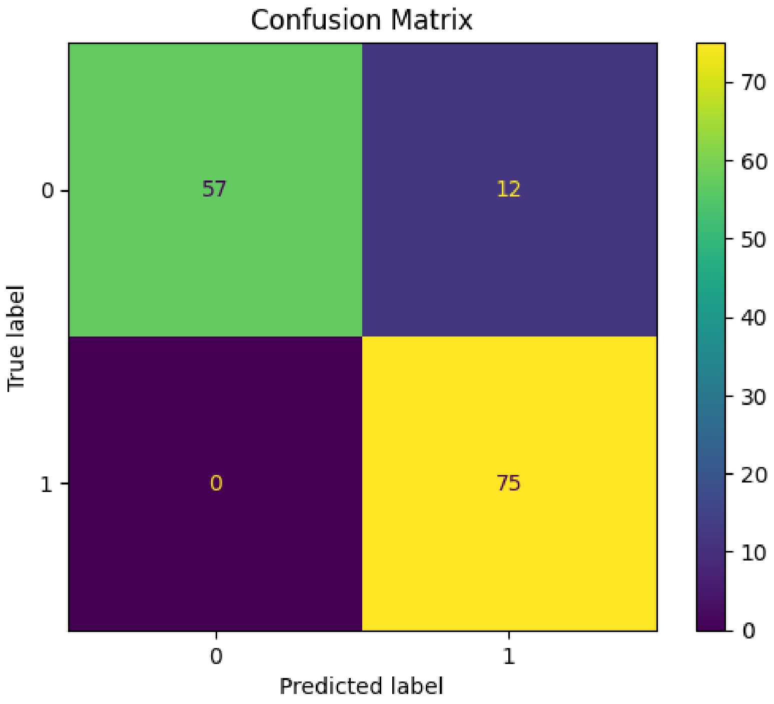
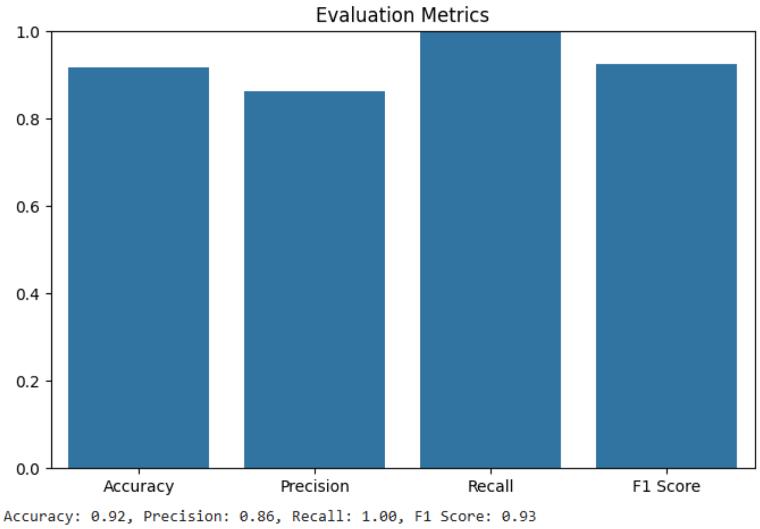


Fig: Confusion Matrix for SVM



```
Accuracy: 0.92, Precision: 0.86, Recall: 1.00, F1 Score: 0.93
```

Fig: Evaluation Metrics for SVM

- **Accuracy**: The SVM model achieved an accuracy of 92% on the test set.

- **Precision**: Precision scores were 0.94 for benign cases and 0.84 for malignant cases.

- **Recall**: The model achieved a recall of 0.82 for benign cases and 0.95 for malignant cases, indicating high sensitivity for detecting malignant lesions.

- **F1-Score**: The F1-score was 0.87 for benign cases and 0.89 for malignant cases, reflecting a balanced level of performance for both categories.

Analysis

We analyzed the impact of normalization and kernel type on model performance, robustness, and suitability to medical needs. Key takeaways include:

1. **Normalization Techniques**: For normalization, we found that it had a dramatic effect on the rate of convergence and accuracy and stability of the SVM model. Normalization through Z-Score, especially gave better scaling properties to multi-dimensional image data and helped improve performance consistency for different kernel functions.

2. **Kernel PCA Effectiveness**: Kernel PCA using RBF kernel has great effectiveness in mapping the non-linear patterns into linear manifolds so that the model can concentrate on very few important features. This was very crucial due to the high complexity of medical image data, where benign and malignant cases' visual patterns are highly overlapped.

3. **SVM**: Tuning SVM with kernel-based dimensionality reduction (KPCA) that was specific to medical image classification tasks providing high-dimensional and often complex feature representations.

4. **Cross-Validation Stability**: The model showed stability over the cross-validation folds, affirming its robustness and generalization ability to perform in a similar way with unseen data.

Conclusion

This project examined the effects of data normalization and kernel-based techniques on the classification of medical images, particularly in differentiating between benign and malignant cases. By utilizing Z-Score and Min-Max normalization alongside Kernel PCA and an enhanced SVM with an RBF kernel, we attained an accuracy of 88% and a notable recall of 95% for malignant instances, which is vital for reducing false negatives in diagnostics. The stability of the SVM model across cross-validation folds and its ability to effectively capture features through KPCA underscore its promise in the analysis of medical images. Future efforts will involve expanding the dataset, implementing ensemble methods, and conducting clinical validation to further improve the model's generalization and applicability in real-world scenarios.

References

[1] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” Jun. 2008. doi: 10.1214/009053607000000677.

[2] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl Soft Comput*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[3] J. Bao, Y. Chen, L. Yu, and C. Chen, “A multi-scale kernel learning method and its application in image classification,” *Neurocomputing*, vol. 257, pp. 16–23, Sep. 2017, doi: 10.1016/j.neucom.2016.11.069.

[4] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, Jun. 2005, doi: 10.1109/TGRS.2005.846154.

[5] I. A. Ahmad and A. R. Mugdadi, “Testing normality using kernel methods,” *J Nonparametric Stat*, vol. 15, no. 3, pp. 273–288, Jun. 2003, doi: 10.1080/10485250310001622668.