Group 44 Interim Report on

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 major 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.

Taken together, the work presented here give 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 from several dermatological studies concentrating on 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, 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.

# Methodology

Thus, the work integrates many techniques of data normalization and bases its systematic evaluation and performance improvement of kernel methods in the classification image. The approach under study includes several principal phases:

1. **Pre-processing**

* **Data Collection:** Employing the database of images of skin lesions, diagnosed all as malignant and benign.
* **Data cleaning:** Removing corrupted images in the dataset of irrelevant or useless output to ensure quality and consistency of data used for training and testing.
* **Data Normalization:** This will be an application of different normalization techniques such as Min-Max Scaling, Z-Score Normalization, or Decimal Scaling to standardize feature scale across the dataset in preventing biasing of kernel calculations.

1. **Kernel Method Selection**

* **SVMs and K-PCA:** These are two main kernel-based techniques that will be focused on, namely SVMs and K-PCA. It is considered because it shows stability and effectiveness in processing high-dimensional, non-linear information structures often suitable for image data.

1. **Modeling and Training**

* **Kernel Function Selection:** Running different kernels and noting varying results will gain a better possibility of capturing complexities in data normalized in this sense.
* **Kernel Optimization**: Using cross-validation and grid search in finding the best values for all kernel techniques to ensure proper and accurate classification can be achieved.
* **Model Training:** The model training will be performed on normalized datasets in which each model was trained with a relatively well-balanced mix of malignant and benign images such that class bias is avoided.

1. **Evaluation**

* **Performance Metrics:** The models will be tested based on accuracy, precision, recall, and F1-score parameters for accurate accuracy of classification of skin lesions.
* **Comparative Analysis**: Comparison of the performance of the models where different normalization techniques have been used to check whether the designed techniques improve the model's accuracy and generalization capabilities.

# 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 Nonparametr Stat*, vol. 15, no. 3, pp. 273–288, Jun. 2003, doi: 10.1080/1048525021000049649.