Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 1, January 2024
https://doi.org/10.47709/cnahpc.v6i1.3600

Submitted : January, 27, 2024
Accepted : February 11, 2024
Published : February 18, 2024

Enhancing Multi-Layer Perceptron Performance with K-Means Clustering

Doughlas Pardede[1)]*, Aulia Ichsan[2)], Sugeng Riyadi[3)]
[1)2)3)]Universitas Deli Sumatera, Indonesia
[1)]doug.pardede@gmail.com, [2)]auliaichsan15@gmail.com, [3)]adhie.ogenk@gmail.com

ABSTRACT

Machine learning plays a crucial role in identifying patterns within data, with classification being a prominent application. This study investigates the use of Multilayer Perceptron (MLP) classification models and explores preprocessing techniques, particularly K-Means clustering, to enhance model performance. Overfitting, a common challenge in MLP models, is addressed through the application of K-Means clustering to streamline data preparation and improve classification accuracy. The study begins with an overview of overfitting in MLP models, highlighting the significance of mitigating this issue. Various techniques for addressing overfitting are reviewed, including regularization, dropout, early stopping, data augmentation, and ensemble methods. Additionally, the complementary role of K-Means clustering in enhancing model performance is emphasized. Preprocessing using K-Means clustering aims to reduce data complexity and prevent overfitting in MLP models. Three datasets - Iris, Wine, and Breast Cancer Wisconsin - are employed to evaluate the performance of K-Means as a preprocessing technique. Results from cross-validation demonstrate significant improvements in accuracy, precision, recall, and F1 scores when employing K-Means clustering compared to models without preprocessing. The findings highlight the efficacy of K-Means clustering in enhancing the discriminative power of MLP classification models by organizing data into clusters based on similarity. These results have practical implications, underlining the importance of appropriate preprocessing techniques in improving classification performance. Future research could explore additional preprocessing methods and their impact on classification accuracy across diverse datasets, advancing the field of machine learning and its applications.

Keywords: Classification; K-Means; MLP; Overfitting; Preprocessing

INTRODUCTION

Machine learning is a widely used technique for identifying patterns in data, particularly in the form of classification (Asad, R., Arooj, S., & Rehman, 2022). In machine learning, there are two common techniques, namely supervised and unsupervised learning, based on the type of data used in the process (Abijono et al., 2021). Supervised learning utilizes data that already have target labels, while unsupervised learning operates on data that do not have targets (Pawluszek-Filipiak & Borkowski, 2020; Yang & Hussain, 2023). Unsupervised learning can assist in identifying patterns or structures that may be challenging to find in labeled data, thus enabling the information to be utilized to enhance the performance of supervised models (Maturo & Verde, 2024).

The Multilayer Perceptron (MLP) algorithm, characterized by an input layer, two or more hidden layers, and an output layer, exhibits resilience to data noise and is capable of solving complex data problems effectively due to its architecture with multiple hidden layers (Pardede & Hayadi, 2023). This algorithm utilizes activation functions, such as ReLu, in the hidden layer to minimize the error values of the output produced by each neuron (Firmansyah & Hayadi, 2022). When dealing with complex data, MLPs are susceptible to overfitting, where the model excessively learns details from the training data and fails to generalize well to test or new data (Dovbnych & Plechawska–Wójcik, 2021). Utilizing data clustering can serve as a helpful strategy in preparing for subsequent modeling steps, aiming to reduce data complexity and prevent overfitting in MLPs (Kolluri et al., 2020).

The K-Means algorithm, a method for non-hierarchical data grouping, endeavors to divide the available data into multiple groups by identifying similar characteristics, thereby efficiently and accurately segregating data with shared attributes from those with differing ones (Tarigan et al., 2023). The K-Means method offers the advantage of being relatively straightforward to implement, capable of managing sizable datasets, and executing the process swiftly (Suwirya et al., 2022). K-Means clustering can serve as preprocessing for other classification methods because it efficiently organizes data into clusters based on similarity, aiding subsequent algorithms in discerning patterns and making accurate predictions (Usman & Stores, 2020). This approach not only streamlines the data preparation process but also has the potential to improve the overall classification performance by uncovering hidden structures within the

* Corresponding author

Submitted : January, 27, 2024
Accepted : February 11, 2024
Published : February 18, 2024

dataset.

This study employs K-Means as a preprocessing model for MLP classification, where data is first grouped based on the number of target categories. The Iris, Wine, and Breast Cancer Wisconsin datasets are utilized to examine the performance of K-Means as a preprocessor across datasets with differing target categories. By utilizing accuracy, precision, recall, and F1 scores obtained from 5-fold, 10-fold, and 20-fold cross-validation evaluations, this research demonstrates how K-Means successfully enhances the performance of MLP classification models across three dataset variations.

LITERATURE REVIEW

Overfitting In MLP

Overfitting is a common challenge encountered in machine learning models, including MLP models, where the model learns noise or irrelevant patterns from the training data, leading to poor generalization on unseen data (Ying, 2019). The complexity of MLP models, including the number of hidden layers and neurons, is a significant contributing factor to overfitting, as heightened complexity increases susceptibility to memorizing noise in the training data rather than discerning meaningful patterns (Rynkiewicz, 2019).

Several researchers have proposed various techniques to mitigate overfitting in MLP models, such as using regularization (Mondal et al., 2020), dropout (Piotrowski et al., 2020), early stopping (Li et al., 2021), data augmentation (Bahtiyar et al., 2022), and ensemble methods (Fayaz et al., 2020). While the effectiveness of techniques such as regularization, dropout, early stopping, data augmentation, and ensemble methods in mitigating overfitting in MLP models has been demonstrated, it is noteworthy that K-Means clustering can also serve as a complementary approach to enhance model performance.

Preprocessing Using K-Means

Overfitting remains a significant challenge in Multilayer Perceptron (MLP) models, necessitating the exploration of innovative preprocessing techniques to enhance model performance (Werner de Vargas et al., 2023). One such approach gaining attention is preprocessing using K-Means clustering, which aims to extract relevant features and reduce data complexity before feeding it into the MLP model (Arvanitidis et al., 2022).

The preprocessing step using K-Means clustering helps in reducing data complexity by grouping similar data points into clusters (Usman & Stores, 2020). This process enables the MLP model to focus on relevant features and avoid memorizing noise or irrelevant patterns present in the data, thus mitigating the risk of overfitting (Walid et al., 2023).

Preprocessing using K-Means clustering complements existing techniques for addressing overfitting in MLP models (Al-Yaseen et al., 2021). While techniques like regularization and dropout directly manipulate the model parameters, K-Means clustering focuses on transforming the input data, providing an additional layer of preprocessing to enhance model robustness and generalization (Andreoni Lopez et al., 2019).

METHOD

Dataset

This study employs three variations of datasets, namely Iris, Wine, and Breast Cancer Wisconsin, obtained from the UCL Machine Learning repository. The selection of these three datasets is based on the differing numbers of target categories in each dataset, with the Breast Cancer Wisconsin dataset having two target categories (Malignant and Benign), Iris having three target categories (Iris Setosa, Iris Versicolor, Iris Virginica), while the Wine dataset has three dataset categories (1, 2, and 3).

Preprocessing

K-Means is employed in preprocessing the data to cluster the data based on the target categories of each dataset. For datasets with two target categories (Breast Cancer Wisconsin), a value of K = 2 is utilized, while for datasets with three target categories (Iris and Wine), a value of K = 3 is used. The clustering results obtained using K-Means Clustering are then utilized as datasets in the MLP classification model.

Model Configuration

The classification model is designed using the MLP algorithm, featuring three hidden layers. Each hidden layer employs 50 neurons, resulting in a configuration of 50-50-50 hidden layers. The activation function utilized in this study is ReLu, which filters the output values from the previous layer within the range of positive values (Margolang

* Corresponding author

et al., 2023). In terms of optimization function, this study utilizes the Adam function, which is capable of efficiently optimizing and resolving a regression problem during the learning process (Firmansyah & Rosnelly, 2023).

Model Evaluation

Cross-validation is employed to assess the performance of the classification model. For each combination of dataset and utilization of K-Means preprocessing, different variations of 5, 10, and 20 folds are utilized. This evaluation yields accuracy, precision, recall, and F1 scores, which are utilized to compare the model performance with and without employing K-Means preprocessing.

RESULT

Performance Without K-Means

By utilizing the previous configuration of the MLP model, evaluation results in terms of accuracy, precision, recall, and F1 scores for the three datasets without preprocessing using K-Means Clustering were obtained through 5-fold, 10-fold, and 20-fold cross-validation, as shown in Table 1.

Table 1. Performance Results Without Preprocessing

Dataset	K-Fold	Accuracy	Precision	Recall	F1
Iris	5	94.7	94.7	94.7	94.7
	10	94.7	94.7	94.7	94.7
	20	95.3	95.3	95.3	95.3
Wine	5	96.6	96.6	96.6	96.6
	10	97.2	97.2	97.2	97.2
	20	96.1	96.1	96.1	96.1
Breast Cancer Winconsin	5	96.8	96.8	96.8	96.8
	10	97.5	97.5	97.5	97.5
	20	96.9	96.9	96.9	96.9

The results depicted in Table 1 illustrate the performance metrics of the MLP model without preprocessing using K-Means Clustering across three different datasets. Across all datasets and folds, high levels of accuracy, precision, recall, and F1 scores were consistently observed, indicating the robustness of the model in accurately classifying the data. Specifically, for the Iris dataset, the accuracy ranged from 94.7% to 95.3%, while for the Wine dataset, it ranged from 96.1% to 97.2%. Similarly, for the Breast Cancer Wisconsin dataset, accuracy ranged from 96.8% to 97.5%. These findings suggest that the MLP model without preprocessing demonstrates strong performance across various datasets and fold configurations.

Performance With K-Means

After employing preprocessing with K-Means Clustering, evaluation results in terms of accuracy, precision, recall, and F1 scores for the three datasets were obtained through 5-fold, 10-fold, and 20-fold cross-validation, as depicted in Table 2.

Table 2Performance Results With Preprocessing

Dataset	K-Fold	Accuracy	Precision	Recall	F1
Iris	5	97.3	97.3	97.3	97.3
	10	97.3	97.3	97.3	97.3
	20	97.3	97.4	97.3	97.3
Wine	5	97.2	97.2	97.2	97.2
	10	97.2	97.2	97.2	97.2
	20	97.8	97.8	97.8	97.8

* Corresponding author

Breast Cancer Winconsin	5	98.7	98.7	98.7	98.7
	10	99.1	99.1	99.1	99.1
	20	98.8	98.8	98.8	98.8

Table 2 presents the evaluation results following preprocessing with K-Means Clustering, showcasing the accuracy, precision, recall, and F1 scores across three datasets through 5-fold, 10-fold, and 20-fold cross-validation. Across all datasets and fold configurations, notably higher values of accuracy, precision, recall, and F1 scores were consistently achieved compared to the results obtained without preprocessing. Specifically, for the Iris dataset, accuracy remained consistently high at 97.3% across all folds, while for the Wine dataset, it ranged from 97.2% to 97.8%. Moreover, for the Breast Cancer Wisconsin dataset, accuracy ranged from 98.7% to 99.1%. These findings underscore the effectiveness of employing preprocessing with K-Means Clustering in enhancing the performance of the MLP model across diverse datasets.

DISCUSSIONS

The results obtained from both the models with and without K-Means preprocessing provide valuable insights into the effectiveness of this technique in enhancing the performance of MLP classification across diverse datasets.

The initial evaluation of the MLP model without K-Means preprocessing yielded promising results across all datasets and fold configurations. Notably, high levels of accuracy, precision, recall, and F1 scores were consistently observed, indicating the robustness of the model in accurately classifying the data. The accuracy of the model ranged from 94.7% to 97.5% across the different datasets, demonstrating its ability to effectively capture the underlying patterns present in the data.

Following preprocessing with K-Means Clustering, the performance of the MLP model exhibited notable improvements across all datasets and fold configurations. The evaluation results showcased significantly higher values of accuracy, precision, recall, and F1 scores compared to the model without preprocessing. Notably, the accuracy of the model increased to a range of 97.2% to 99.1% across the datasets, indicating a substantial enhancement in classification performance.

Comparing the performance metrics of the model with and without K-Means preprocessing reveals the significant impact of this technique in improving the overall classification accuracy. The consistent increase in accuracy, precision, recall, and F1 scores across all datasets underscores the effectiveness of K-Means Clustering in enhancing the discriminative power of the MLP model.

The findings of this study demonstrate the effectiveness of preprocessing with K-Means Clustering in improving the performance of MLP classification models. By effectively capturing the underlying structure of the data, K-Means preprocessing enables the model to achieve higher levels of accuracy and precision in classifying the datasets. These results highlight the importance of employing appropriate preprocessing techniques to enhance the performance of machine learning models in real-world applications. Further research could explore the potential of other preprocessing methods and their impact on classification performance across different types of datasets.

CONCLUSION

This study explored the effectiveness of preprocessing using K-Means Clustering in enhancing the performance of Multilayer Perceptron (MLP) classification models across diverse datasets. Overfitting, a common challenge in MLP models, was addressed through the innovative application of K-Means clustering to streamline data preparation and improve classification accuracy. The investigation began with a comprehensive overview of overfitting in MLP models, emphasizing the importance of addressing this issue to ensure robust performance in classification tasks. Various techniques for mitigating overfitting were reviewed, including regularization, dropout, early stopping, data augmentation, and ensemble methods. While these techniques have shown promise, the literature review highlighted the complementary role of K-Means clustering in enhancing model performance. Preprocessing using K-Means clustering was then examined as a method to reduce data complexity and prevent overfitting in MLP models. The study employed three datasets, namely Iris, Wine, and Breast Cancer Wisconsin, to evaluate the performance of K-Means as a preprocessing technique. Results obtained through cross-validation demonstrated significant improvements in accuracy, precision, recall, and F1 scores when employing K-Means clustering compared to models without preprocessing. The findings underscored the effectiveness of K-Means clustering in enhancing the discriminative power of MLP classification models. By organizing data into clusters based on similarity, K-Means preprocessing enabled the model to focus on relevant features and avoid overfitting. Notably, the accuracy, precision,

Submitted : January, 27, 2024
Accepted : February 11, 2024
Published : February 18, 2024

recall, and F1 scores consistently showed substantial enhancements when K-Means preprocessing was applied, highlighting its efficacy in improving model performance. These results have implications for real-world applications, emphasizing the importance of appropriate preprocessing techniques in improving classification performance. In conclusion, this study contributes to the growing body of research on preprocessing methods for enhancing machine learning model performance. Future research could further explore the potential of other preprocessing techniques and their impact on classification accuracy across various datasets, ultimately advancing the field of machine learning and its practical applications.

REFERENCES

Abijono, H., Santoso, P., & Anggreini, N. L. (2021). Supervised Learning and Unsupervised Learning Algorithms in Data Processing. *Jurnal Teknologi Terapan: G-Tech*, *4*(2), 315–318. https://doi.org/10.33379/gtech.v4i2.635

Al-Yaseen, W. L., Jehad, A., Abed, Q. A., & Idrees, A. K. (2021). The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer. *International Journal of Intelligent Engineering and Systems*, *14*(2), 190. https://doi.org/10.22266/ijies2021.0430.17

Andreoni Lopez, M., Mattos, D. M. F., Duarte, O. C. M. B., & Pujolle, G. (2019). A fast unsupervised preprocessing method for network monitoring. *Annales Des Telecommunications/Annals of Telecommunications*, *74*(3–4), 139–155. https://doi.org/10.1007/s12243-018-0663-2

Arvanitidis, A. I., Bargiotas, D., Daskalopulu, A., Kontogiannis, D., Panapakidis, I. P., & Tsoukalas, L. H. (2022). Clustering Informed MLP Models for Fast and Accurate Short-Term Load Forecasting. *Energies*, *15*(4), 1–14. https://doi.org/10.3390/en15041295

Asad, R., Arooj, S., & Rehman, S. U. (2022). Study of Educational Data Mining Approaches for Student Performance Analysis. *Technical Journal*, *27*(1), 68-81. https://www.researchgate.net/publication/362762123_Study_of_Educational_Data_Mining_Approaches_for_Student_Performance_Analysis

Bahtiyar, H., Soydaner, D., & Yüksel, E. (2022). Application of multilayer perceptron with data augmentation in nuclear physics. *Applied Soft Computing*, *128*(August). https://doi.org/10.1016/j.asoc.2022.109470

Dovbnych, M., & Plechawska–Wójcik, M. (2021). A comparison of conventional and deep learning methods of image classification. *Journal of Computer Sciences Institute*, *21*(September), 303–308. https://doi.org/10.35784/jcsi.2727

Fayaz, M., Khan, A., Rahman, J. U., Alharbi, A., Uddin, M. I., & Alouffi, B. (2020). Ensemble machine learning model for classification of spam product reviews. *Complexity*, *2020*. https://doi.org/10.1155/2020/8857570

Firmansyah, I., & Hayadi, B. H. (2022). Komparasi Fungsi Aktivasi Relu Dan Tanh Pada Multilayer Perceptron. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(2), 200. https://doi.org/10.26798/jiko.v6i2.600

Firmansyah, I., & Rosnelly, R. (2023). Inception-V3 Versus VGG-16 : in Rice Classification Using Multilayer Perceptron. *2nd International Conference on Information Science and Technology Innovatin (ICoSTEC)*, *2(1)*, 1–5. https://prosiding-icostec.respati.ac.id/index.php/icostec/article/view/24

Kolluri, J., Kotte, V. K., Phridviraj, M. S. B., & Razia, S. (2020). Reducing Overfitting Problem in Machine Learning Using Novel L1/4 Regularization Method. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, *June*, 934–938. https://doi.org/10.1109/ICOEI48184.2020.9142992

Li, T., Zhuang, Z., Liang, H., Peng, L., Wang, H., & Sun, J. (2021). Self-Validation: Early Stopping for Single-Instance Deep Generative Priors. *32nd British Machine Vision Conference, BMVC 2021*, 1–14.

Margolang, K. F., Riyadi, S., Rosnelly, R., & Wanayumini. (2023). Pengenalan Masker Wajah Menggunakan VGG-16 dan Multilayer Perceptron. *Jurnal Telematika*, *17*(2), 80–87.

Maturo, F., & Verde, R. (2024). Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers. *Computational Statistics*, *39*(1), 239–270. https://doi.org/10.1007/s00180-022-01259-8

Mondal, R., Dey, P., Sharma, G., & Pal, T. (2020). Regularizing Multilayer Perceptron for Generalization Using KL-Divergence. *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*. https://doi.org/10.1109/ICCSEA49143.2020.9132891

Pardede, D., & Hayadi, B. H. (2023). Klasifikasi Sentimen Terhadap Gelaran MotoGP Mandalika 2022 Menggunakan Machine Learning. *Jurnal TRANSFORMATIKA*, *20*(2), 42–50.

Pawluszek-Filipiak, K., & Borkowski, A. (2020). On the importance of train-test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing*, *12*(18), 0–32. https://doi.org/10.3390/rs12183054

* Corresponding author

Submitted : January, 27, 2024
Accepted : February 11, 2024
Published : February 18, 2024

Piotrowski, A. P., Napiorkowski, J. J., & Piotrowska, A. E. (2020). Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling. *Earth-Science Reviews*, *201*(August 2019), 103076. https://doi.org/10.1016/j.earscirev.2019.103076

Rynkiewicz, J. (2019). On overfitting of multilayer perceptrons for classification. *ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, *April*, 257–262.

Suwirya, I. P., Candiasa, I. M., & Dantes, G. R. (2022). Evaluation of ATM Location Placement Using the K-Means Clustering in BNI Denpasar Regional Office. *Journal of Computer Networks, Architecture and High Performance Computing*, *4*(2), 158–168. https://doi.org/10.47709/cnahpc.v4i2.1580

Tarigan, N. M. B., Tarigan, S. E. B., & Simatupang, A. P. (2023). Implementation of Data Mining in Grouping Data of the Poor Using the K-Means Method. *Journal of Computer Networks, Architecture and High Performance Computing*, *5*(2), 599–611. https://doi.org/10.47709/cnahpc.v5i2.2625

Usman, D., & Stores, F. S. (2020). On Some Data Pre-processing Techniques for K-Means Clustering Algorithm. *Journal of Physics: Conference Series*, *1489*(1). https://doi.org/10.1088/1742-6596/1489/1/012029

Walid, M., Sahbaniya, N. L. N., Hozairi, H., Baskoro, F., & Wijaya, A. Y. (2023). K-Means Clustering and Multilayer Perceptron for Categorizing Student Business Groups. *Knowledge Engineering and Data Science*, *6*(1), 69. https://doi.org/10.17977/um018v6i12023p69-78

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, *65*(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8

Yang, M. S., & Hussain, I. (2023). Unsupervised Multi-View K-Means Clustering Algorithm. *IEEE Access*, *11*, 13574–13593. https://doi.org/10.1109/ACCESS.2023.3243133

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2). https://doi.org/10.1088/1742-6596/1168/2/022022