International Journal of Science and Technology Research Archive

ISSN: 0799-6632 (Online)

Journal homepage: https://sciresjournals.com/ijstra/

(Research Article)

K-Means Clustering of the Iris dataset using the Perl language

Diptarshi Mitra *

Department of Computer Applications, Institute of Hotel & Restaurant Management, Sonarpur (PIN: 700150), India

International Journal of Science and Technology Research Archive, 2024, 06(01), 089–091

Publication history: Received on 04 March 2024; revised on 14 March 2024; accepted on 15 March 2024

Article DOI: https://doi.org/10.53771/ijstra.2024.6.1.0031

Abstract

Generally, the Perl programming language is not used in the domains of Machine Learning and Data Science. This study tries to find out whether Perl can be used for writing codes involving Machine Learning and Data Science. Here, it has been observed that Perl can be used to successfully implement the K-Means Clustering algorithm from scratch. Thus, Perl may be used for solving problems by applying Machine Learning and Data Science.

Keywords: Perl; K-Means Clustering; Iris Dataset; Machine Learning; Data Science

1. Introduction

Python and R programming languages are generally used nowadays for applying Machine Learning and Data Science to various fields. Besides, there are a number of languages, like MATLAB (MATrix LABoratory), Java etc., which are also often applied to these two domains (i.e., Data Science and Machine Learning). However, there are lots of programming languages in the realm of Computer Science. And, some languages are normally not employed for writing codes involving Data Science and Machine Learning. One such language is Perl (Practical Extraction and Reporting Language). It is a general purpose, high level, interpreted and dynamic language. This study has tried to find out whether Perl can be utilized for writing programs involving Data Science and Machine Learning. If Perl can be applied successfully to these domains, then Perl programmers can extend their horizon, and contribute extensively to problem solving by employing Data Science and Machine Learning.

Before writing this article, a literature survey was conducted to have a look at similar studies by other scientists and researchers; the outcome is as follows.

Baiocchi extensively demonstrated the applicability of Perl in the field of Statistics (Baiocchi, 2004).

Marino et al. used Perl to perform various pre-processing and post-processing tasks pertaining to Compressive Big Data Analytics (CBDA) 2.0, a technique (developed by Marino et al.) for handling Big Data (Marino et al., 2020).

Sahu and Mehtre worked on network intrusion detection, using the Decision Tree (J48) algorithm and Kyoto 2006+ dataset, and utilized Perl for extracting features from the said dataset (Sahu & Mehtre, 2015).

However, so far, no journal/conference article has been found, in which Perl has been used for directly implementing a Machine Learning technique.

* Corresponding author: Diptarshi Mitra

2. Methodology

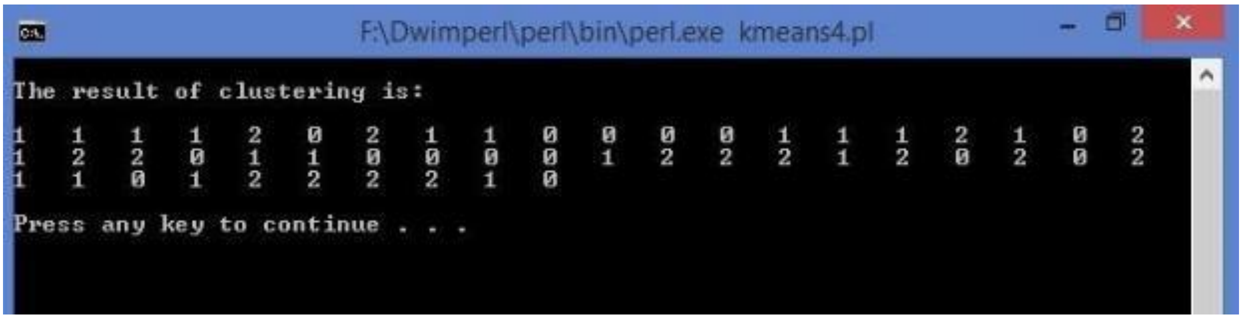
In this work, Perl language has been used to apply the K-Means Clustering algorithm on the Iris dataset. Here, the K-Means Clustering method has been implemented from scratch.

K-Means Clustering technique is an unsupervised Machine Learning algorithm, used to divide a set of data items into a number of groups (or clusters). First, k means or cluster centroids are randomly chosen (k=number of clusters into which the data items need to be grouped). Then, the Euclidean distance between each cluster centroid and a data item, is calculated; the data item is assigned to that cluster for which the distance between the said item and the corresponding cluster centroid, is the minimum. Subsequently, the coordinates of the centroid, pertaining to the cluster in which the data item has been included, are updated. Afterwards, the Euclidean distance between each cluster centroid and the next data item, is computed, and the data item is included in that cluster for which this distance is the minimum. And then, the coordinates of the corresponding cluster centroid are updated. This process is repeated for a fixed number of data items (which may be called the training dataset). Then, on the basis of the cluster centroids (with updated coordinates), a number of data items (which comprise the test dataset) are clustered. While clustering the items of the test dataset, the coordinates of the centroids are not updated.

The Iris dataset contains data on the sepal length, sepal width, petal length, and petal width, of 150 Iris flowers belonging to three categories viz., Setosa, Versicolour and Virginica (Kaggle, 2024). In this study, the rows of the Iris dataset have been randomly shuffled, and the column names have been removed, before applying the K-Means algorithm (for performing these two tasks, Perl programming has not been employed). Also, it may be noted that the training dataset consists of 100 data items, and the test dataset contains the remaining 50.

3. Result and Discussion

Fig-1 shows the outcome of clustering the test dataset by employing the K-Means algorithm, implemented through Perl programming.



```
The result of clustering is:

1       1       1       1       2       0       2       1       1       0       0       0       0       1       1       1       2       1       0       2
1       2       2       0       1       1       0       0       0       0       1       2       2       2       1       2       0       2       0       2
1       1       0       1       2       2       2       2       1       0

Press any key to continue . . .
```

Figure 1 Output of the K-Means Algorithm

After checking the original (Iris) dataset, it has been found that the accuracy of clustering is 94%. Thus, the K-Means Clustering algorithm has been successfully implemented by utilizing Perl.

So, it is expected that Perl can be used for successfully implementing other Machine Learning algorithms as well. Thus, Perl programmers can think of writing programs involving Machine Learning and Data Science.

4. Conclusion

An unsupervised Machine Learning algorithm viz., the K-Means Clustering method, has been successfully implemented using Perl. Thus, Perl may be considered for problem solving by employing Machine Learning and Data Science.

Compliance with ethical standards

Acknowledgements

I would like to thank all the academic and administrative staff members of the Institute of Hotel & Restaurant Management, for their help, cooperation and support.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Baiocchi, G. (2004). Using Perl for Statistics: Data Processing and Statistical Computing. *Journal of Statistical Software*, *11*(1).

[2] Kaggle. (2024). Iris Flower Dataset. Retrieved from https://www.kaggle.com/datasets/arshid/iris-flower-dataset.

[3] Marino, S., Zhao, Y., Zhou, N., Zhou, Y., Toga, A. W., Zhao, L., … Dinov, I. D. (2020). Compressive Big Data Analytics: An Ensemble Meta-Algorithm for High-Dimensional Multisource Datasets. *PLoS ONE*, *15*(8).

[4] Sahu, S., & Mehtre, B. M. (2015). Network Intrusion Detection System Using J48 Decision Tree. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2023–2026). Kochi (India).