

Comparative Analysis of Electricity Consumption at Home through a Silhouette-score prospective

Hyun Wong Choi, Nawab Muhammad Faseeh Qureshi and Dong Ryeol Shin

Sungkyunkwan University, South Korea.

pooh0216@g.skku.edu

Abstract— Machine learning is a state-of-the-art sub-project of artificial intelligence, that is been evolved for finding large-scale intelligent analytics in the distributed computing environment. In this paper, we perform comparative analytics onto dataset collected for the electricity usage of home based on the K-mean clustering algorithm using comparison to silhouette score with a ratio of 1/8 dataset. The performance evaluation shows that, the comparison index is similar in numbers of silhouette score even if dataset is smaller than before.

Keywords— K-means clustering, Machine learning, Unsupervised Learning, Silhouette score.

I. INTRODUCTION

Machine learning is an analysing mechanism that fetches and identifies the matching patterns from existing datasets for newer result formations. This paper discusses comparative analytics related to unsupervised learning algorithms, at which we compare the K-mean clustering result with a ratio of half dataset to Silhouette_score results. We performed analysis and came to a conclusion that Davis-Boulden index is not working smoothly in the sci-kit learning, so performed a check analysis for Caliski-Harabasz Index and Silhouette score along with and Davis – Boulden index and compared results to each of them so to learn that when we reduce the dataset to a mentioned proportion, the resultant dataset shows half score than the traditional dataset score.

II. PREVIOUS WORKS

Machine learning is a field of artificial intelligence, that is used to develop algorithms and techniques that enable computers to learn [1]. It is used to train the computer to distinguish whether e-mails received are spam or not, and there are various applications such as data classification, associated rule identification, and character recognition, which comply to the standard machine learning perspectives.

It includes a series of processes, in which a computer finds its own patterns, creates a new classification system, analyzes the data, and produces meaningful results. The successful prediction occurs with the increase in probability and decrease in the error issues. Machine learning enables sort out the issues with various iterative learning [2]. Among them, the supervised learning is highly related to summarizing the learning methods for re-enforcement mechanisms [3].

Clustering is a process of mining the dataset by defining a cluster of data that considers the characteristics of input and finds a representative method to point out the data group. In this way, a cluster is a group of relevant data elements with similar characteristics. If the functions are not same, the ingredients belong to contrast clusters [3]. Clustering is an unsupervised learning without accuracy in answers. In the same way, the objects having same information are grouped together for similar elements. However, the classification is a way related to supervised learning. When you perform classification operations, the system will learn to predict the dependent variable (Y) with the independent variable (X) of the data [4].

Scikit-learn is the machine learning platform in the middle range of superficial broad python module this package high-level language can us easily high-level documentation and proper API suggested. Using BSD license as academic or commercially use it. Source-code, documentation is downloaded from websites [10]. Supervised learning, Unsupervised Learning is the many problems is inserted in the Scikit-learn, Generalized Models, Linear and Quadratic Decruitment Analysis, Kernel Ridged regression, Support Vector machine, Stochastic Gradient Decent model's solution also inserted in the Scikit-learn.

III. PROPOSED APPROACH

K-means algorithm is one of the clustering methods for divided, divided is giving the data among the many partitions. For example, receive data object n, divided data is input data divided K ($\leq n$) data, each group consisting of cluster below equation is the at K-means algorithm when cluster consists of algorithms using cost function use it [11]

$$\operatorname{argmin} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2$$

In other words, one of the data objects divided by the K group. Currently divided similarity is (dissimilarity with reducing the cost function about it. And from this theory each object similarity increase, different group similarity will decrease. [12] K-means algorithm is each centroid and in each group's data object times' summation, from this function result, the data object group updated clustering progressed. [5]

Silhouette score is the easy way to in data I each data cluster in data's definition an (i) each data is not clustered inner and data's definition b(i) silhouette score s(i) is equal to calculate that

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}}$$

From this calculate s(i) is equal to that function

$$-1 \leq s(i) \leq 1$$

S(i) is the close to 1 is the data I is the correct cluster to each thing, close to -1 cannot distribute cluster is distributed, from this paper machine Using the machine learning library scikit-learn in the house hold power consumption clustering [7]. Household power consumption from the dataset Download from University California Irvine Machine Learning Dataset Repository [8] and then use it, this dataset is via delimiter is divided. Global active power, Global Reactive power, Voltage, Global_intensity is divided. Global Active_power and Global Reactive power the X, Y axis experiment it. Python library is Anaconda3 K-means algorithm's key point is using Data keep K clusters, reduce cluster's distance, K-means algorithms input data put the labels. figure 1 is the before check Calinski-Harabasz Index and Silhouette_score execute K-means algorithm's result. Figure 1 to Figure 11 are 1/8 dataset k-means clustering result for House hold power consumption from UC Irvine Repository and reduce the dataset 1/8 times from original UCI machine learning data repository.

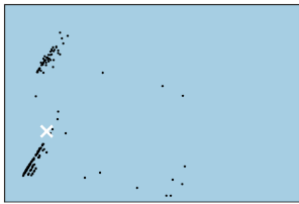


Figure 1. 1/8 dataset cluster K=1

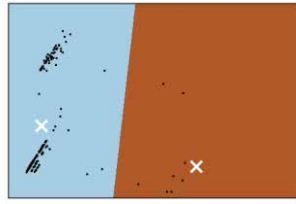


Figure2. 1/8 dataset cluster K=2

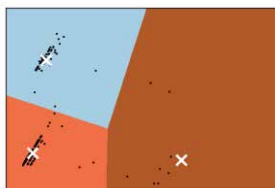


Figure3. 1/8 dataset cluster K=3

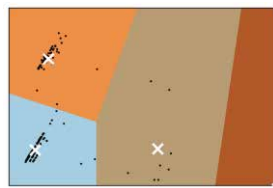


Figure4. 1/8 dataset cluster K=4

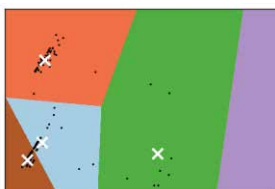


Figure5. 1/8 dataset cluster K=5



Figure6. 1/8 dataset cluster K=6



Figure7. 1/8 dataset cluster K=7

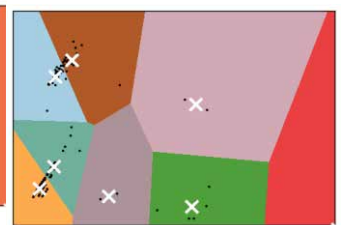


Figure8. 1/8 dataset cluster K=8



Figure9. 1/8 dataset cluster K=9



Figure10. 1/8 dataset cluster K=10



Figure11. 1/8 dataset cluster K=11

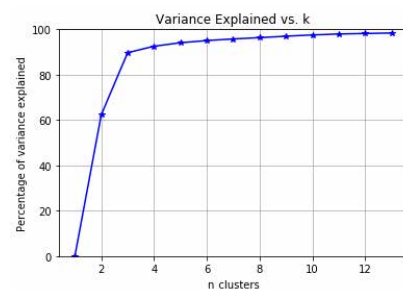


Figure 12. Silhouette score according to change of cluster number.

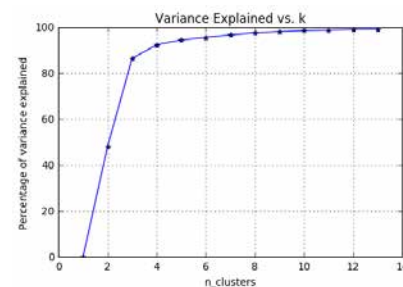


Figure 13. 1/8 dataset Silhouette score according to change of cluster number.

From K-means algorithms calculate proper cluster things is very important, from the data, estimate Silhouette_score, the result is $K = 7$ each cluster centroid and data prices silhouette score is 0.799 is the optimal score. Even if dataset is so small but the 1/8 datasets $K = 7$ each cluster centroid and data prices silhouette score 0.810 is the optimal score. From this K-means algorithm cluster 7th, (all dataset , 1/8 dataset) each group's centroid and each centroid distance will be an optimal value. From this result, the dataset is decrease but the K-means clustering 's class vector space. Its optimal cluster is same situation with before original Dataset Household power consumption rate via clustering.

IV. CONCLUSIONS

From the paper, Household power consumption via k-means clustering, Used library which is sci-kit learn, Anaconda 3 open-source personally can easily follow it and because using BSD License to real works don't have difficulties to that. From this result even if reduce the dataset 1/8 but the silhouette score and all the clustering result is same with before. But the population will increase it can show the clearer result for the classification and vector space. Large dataset to small dataset is clear to show to the specific result for the Silhouette score but opposite site is not clearly allowed. Because 4-dimension vector dataset. From the experiment reduce the estimated time if received huge dataset from analysis.

REFERENCES

- [1] https://en.wikipedia.org/wiki/K-means_clustering
- [2] https://en.wikipedia.org/wiki/Cluster_analysis
- [3] [https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering))
- [4] <https://github.com/sarguido>.
- [5] <http://archive.ics.uci.edu/ml/datasets.html>.
- [6] <http://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index>
- [7] <http://scikit-learn.org/stable/>.
- [8] T. Calinski and J. Harabasz, 1974. "A dendrite method for cluster analysis". *Communications in Statistics*
- [9] Kanungo, Tapas et al. "An Efficient k-Means Clustering Algorithm: Analysis and Implementation." *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (2002): 881-892.
- [10] David, and Sergei Vassilvitskii, "k-means++: The advantages of careful seeding" *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007): 1027-1035
- [11] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In *ICML* (Vol. 1, pp. 577-584).



Hyun Wong Choi was born in Republic of Korea, Feb 16, 1988. Hyun Wong Choi is master's degree from Sungkyunkwan University in Korea. His main research interests include Machine learning, IoT, Software Defined Network.



NAWAB MUHAMMAD FASEEH QURESHI is an Assistant Professor in Department of Computer Education, Sungkyunkwan University, Seoul, South Korea. He received Ph.D. degree in Computer Engineering from Sungkyunkwan University, South Korea with the funding support of SAMSUNG scholarship and was awarded with the Superior Research Award from the College of Information and Communication Engineering on account of his research contributions and performance during Ph.D. He is part of editorial and reviewer of various prestigious journals *Future Generation Computer Systems*, *Transactions on Emerging Telecommunications Technologies (ETT)*, *Wireless Personal Communications*, *KSII Transactions on Internet and Information Systems*, *Journal of Supercomputing*, *IEEE Communications Magazine*, *IEEE Transactions on Industrial Informatics* and *IEEE Access*. He is also reviewer of various top-tier conferences such as *IEEE Globecom2018* and *IEEE PIMRC 2017*. He has been in TCP with *IWWCN2017*, *CSA2017*, *IMTIC18* and *WCSN2017* conferences and performed as session chairs with *ICGCET Denmark* and soon with *RTCSE19 USA*. He has facilitated institutes with *Webinars on Big data analysis (SZABIST Larkana)* and *Modern Technology Convergence (MSF South Korea)*. He was also invited for a keynote talk on convergence with modern technologies in *IEEE Pakistan Seminar* conducted in *MUET*. He is an active member of *IEEE*, *ACM*, *KSII* and *IEICE*. His research interests include big data analytics, context-aware data processing of the Internet of Things, and cloud computing.



DONG RYEOL SHIN received the B.S. degree in electrical engineering from the Sungkyunkwan University in 1980, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology in 1982, and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, USA, in 1992. From 1992 to 1994, he was with Samsung Data Systems, South Korea, where he was involved in the research of intelligent transportation systems. Since 1994, he has been with the Department of Computer Science and Engineering, Sungkyunkwan University, where he is currently a Full Professor with the Network Research Group. His current research interests lie in the areas of mobile network, ubiquitous computing, cloud computing, and bioinformatics. He is actively involved in the security of vehicular area networks, and the implementation and analysis of Big Data platform, applicable to 3-D image processing of robotic arms.