Analysis of Electricity Consumption at Home Using K-means Clustering Algorithm

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

Sungkyunkwan University, South Korea.

pooh0216@g.skku.edu

Abstract— **Machine learning is a modern field that has emerged as a new tool for data analytics in the distributed computing environment. There are several aspects, at which, machine learning has improved the processing capacity along with effectiveness of analysis. In this paper, the electricity usage of home is analyzed through K-means clustering algorithm for obtaining the optimal home usage electricity data points. The Davis Boulden Index and Silhouette_score finds the detailed optimal number of clusters in the K-means algorithm and present the application scenario of the machine learning clustering analytics.**

Keywords— **K-means clustering, Machine learning, Unsupervised Learning, Davis Boulden Index.**

I. INTRODUCTION

Machine learning is a sub-project of artificial intelligence, that is used to develop algorithms and techniques for enabling the computers to learn [1]. It is used to train the computer for various aspects such as (i) distinguish whether e-mails received are spam or not, (ii) data classification application, (iii) association rule identification, and (iv) character recognition.

Machine learning includes a series of processes, in which a computer lookup for (i) similar patterns, (ii) generate a novel classification system, (iii) data analytics, and (iv) producing meaningful results. It is a kind of artificial intelligence, that can be predicted based on the result, if it is supported only by analytics algorithms. Machine learning is a step-by-step evolution process that leads from big data analytics to predict future actions towards making decisions on its own through past learned results. The key issues for processing a successful prediction model remains to be within increasing the probability and reducing the error and the said problems are resolved through enabling numerous iterative learnings [2].

At the heart of machine learning are Representation and Generalization, where expression is an evaluation of data and generalization is processing of future data. Unsupervised learning is a type of machine learning that is used primarily to determine how data is organized. Unlike Supervised Learning or Reinforcement Learning, this method does not give a target value for input values [3].

Autonomous learning is closely related to the density estimation of statistics. These autonomous learning can summarize and describe the main characteristics of the data. An example of autonomous learning is clustering. In this paper, we use the K-means algorithm to measure the optimal number of clusters based on the Calinski-Harabasz Index and Silhouette_score, Davis-Boulden index and then apply it to household electricity consumption analysis.

II. PREVIOUS WORKS

1. Machine Learning

Machine learning is like data mining, but it is different in predicting data based on learned attributes, mainly through training data. In addition to the three techniques, Unsupervised learning, Supervised Learning or Reinforcement Learning, various types of machine learning techniques such as Semi-Supervised Learning and Deep Learning algorithms are developed Has been used.

2. Clustering

Clustering is a method of data mining by defining a cluster of data considering the characteristics of given data and finding a representative point that can represent the data group. A cluster is a group of data with similar characteristics. If the characteristics of the data are different, they must belong to different clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including pattern recognition, information retrieval, machine learning, and computer graphics [3].

(1) Maximizing inter-cluster variance

(2) Minimizing the inner-cluster variance

Note, however, that clustering should be distinguished from Classification. Clustering is unsupervised learning without correct answers. In other words, we group similar objects without group information of each object. Classification, on the other hand, is supervised learning. When you carry out classification tasks, you will learn to predict the dependent variable (Y) with the independent variable (X) of the data [4].

3. Community Feasibility Assessment

Since clustering tasks are not correct, they cannot be evaluated as indicators, such as simple accuracy, as in a typical machine learning algorithm. As you can see in the example below, it is not easy to find the optimal number of clusters

without the correct answers. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include group with small distances between cluster members, dense areas of data space,

4. Scikit-learn

In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. If each sample is more than a single number and for instance. A multi-dimensional entry, it is said to have several attributes or features.

Supervised learning, in which the data comes with additional attributes that we want to predict this problem can be either. Classification : samples belong to two or more classes and we want to learn from already labelled data how to predict the class of unlabeled data. An exmaple of a classification problem would be handwritten digit recognition, in which the aim is to assign each input vector to one of a finite number of discrete categories. Another way to think of classification is as a discrete(as opposed to continuous) form of supervised learning where one has a limited number of categories and for each of n samples provided. One if to try to label them with the correct category or class.

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]

How to be well to be clustering inner way is Caliski-Harabasz Index, Davies-Bouldin index, Dunn index, Silhouette score. In this paper. Evaluate via Clainiski-Harabasz Index and silhouette score evaluate it.

From the Cluster Calinski-Harabasz Index s I the clusters distributed average and cluster distributed ratio will give it to you.

$$s(k) = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-k}{k-1}$$

For this Bk is the distributed matrix from each group Wk is the cluster distributed defined.

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q} n_q (C_q - c)(C_q - c)^T$$

N is the number of Data, Cq data group in Cq, Cq is the cluster q's centroid, c is the E of the Centroid, NQ is the number of data number in cluster_q

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

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

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

$$-1 \leq \mathrm{s(i)} \leq 1$$

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 k-means clustering result for House Hold power consumption from UC Irvine repository.

Figure 1. Clustering result at K = 1

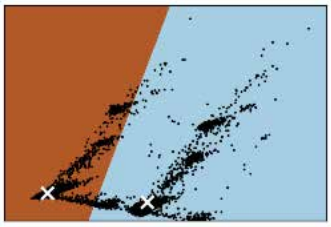

Figure2. Clustering result at K=2

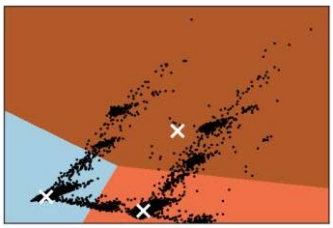

Figure 3. Clustering result at K = 3

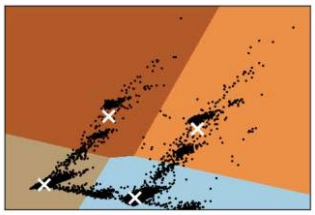

Figure4. Clustering result at K=4

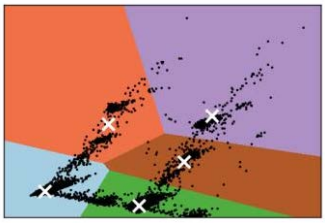

Figure 5. Clustering result at K = 5

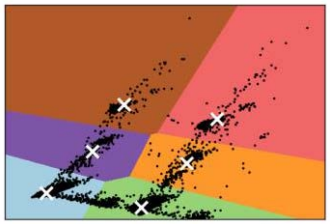

Figure 6. Clustering result at K=6

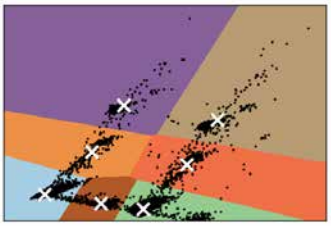

Figure 7. Clustering result at K = 7

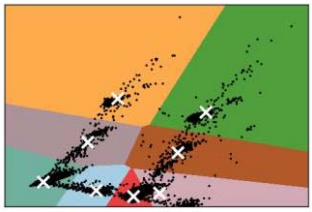

Figure 8. Clustering result at K=8

Figure 9. Clustering result at K = 9

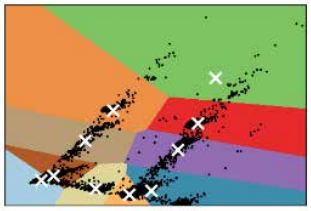

Figure 10. Clustering result at K=10

After all, reduce each cluster's distance calculate each cluster's Calinski-Harabasz Index, increasing clusters' Calinski-Harabasz Index will decrease with K ratio is too law estimate K this cluster partition will one more or not electric consumption rate is very important. This one is the most important fact.

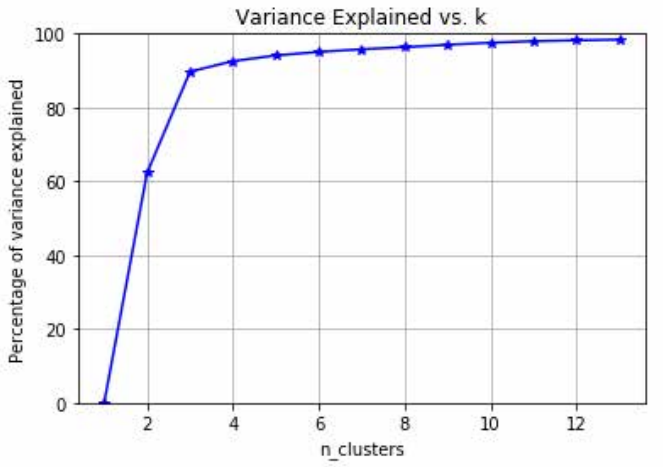

Figure 11. Shiloutette score according to change of cluster number.

Equal with Caliski-Harabasz Index estimation, calculate Silhouette_score. The cluster will increase Silhouette_score will decreases with K distributed, a low factor with optimal K represented.

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 are 0.799 is the optimal score. From the formal Caliski-Harabasz Index results are 560.3999 is the optimal result. Using this k-means algorithm the fact is figure 11.

From this K-means algorithm cluster $7^{th,}$ each group's centroid and each centroid distance will be an optimal value. From this result, each Centroid can divide. Household power consumption rate via clustering.

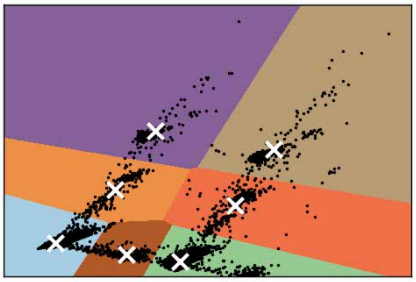

Figure 12: Clustering result at K=7

Davies-Bouldin index

If the ground truth labels are not known, the Davies-Bouldin index (sklearn. Metrixdavis Boulden)

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then the Davis-Bouldin Index is defined as

$$\mathrm{DB} = \frac{1}{k}\sum i = 1^{k} \max_{i \neq j} R_{ij}$$

The zero is the lowest score a possible. Score. Values closer to zero indicate a better partition. But the problem is this algorithm do not attach it in the Scikit-learn library and only explain it in the document page but cannot experiment easily.

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. Not only the K-means algorithm, PCA Algorithms, but also SVM algorithm etc other machine learning algorithms clustering can also do it. From this result, in real life household power consumptions diverse analytics. And electricity transformer, Transmission power can management period can estimate it. And each data using electricity consumption. It can be used for progressive taxation, regional to regional demand forecasting, maintenance of power plants and facilities. Can do it. In the Gas company can estimate via k-means algorithms and also can estimate about the gas consumption rate to via K-means clustering and index.

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)
[4] https://github.com/sarguido.
[5] http://archive.ics.uci.edu/ml/datasets.html.
[6] http://scikit-learn.org/stable/modules/clustering.html#calinski-harabaz-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).
[12] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.
[13] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), 881-892.
[14] Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.
[15] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.
[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.
[17] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Layton, R. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.
[18] Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, *8*, 14.
[19] Fabian, P., Gaël, V., Alexandre, G., Vincent, M., Bertrand, T., Olivier, G., ... & Alexandre, P. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.
[20] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

HyunWong Choi, was born in Republic of Korea, Feb 16, 1988. Hyun Wong Choi is Master 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 Edu cation, 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 mod-ern technologies in IEEE Pakistan Seminar conduced 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.