ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse51942020.pdf
https://doi.org/10.30534/ijatcse/2020/51942020

A Comparative Approach to Evaluate Different CVIs using Grid K-Means and Improved K-Means Clustering

Hutashan Vishal Bhagat[1], Dr. Manminder Singh[2]
[1]Department of Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India, hutashan20@gmail.com
[2]Department of Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, Longowal, Punjab, India, manminderldh@gmail.com

ABSTRACT

Talk about the era we are living in today, is not limited to just include the living creatures but also the most important aspect upon which these living creatures are now more rely i.e., "Information". There is an abundance of information against every aspect available on this planet. So, the term "Data" is too concise to encapsulate this "Information". Hence, the era of Big Data come into existence. Data is now big enough (in terms of volume, variety, value, veracity, velocity) that without proper techniques and methods it is not possible to frame a definite set of knowledgeable data. A need to mine this deep ocean of information to get knowledgeable data results in the various Data Mining techniques. In this paper, a reflection of all the major data analysis techniques and how the traditional models are replaced by the new emerging technologies based on machine learning or deep learning is presented. Partition-based clustering is the most commonly used technique of unsupervised learning; in this paper, Improved K-Means and Grid K-Means algorithms are used to form clusters for four publically available datasets. The effectiveness of these clusters is evaluated using seven different Cluster Validity Indexes. Results show that VCVI-index and BVCI-index outperform among all other CVIs.

Key words: Cluster Validity Index, CVIs, Grid K-Means, Improved K-Means, Machine Learning Techniques, Supervised Learning

1. INTRODUCTION

World is today linked with several heterogeneous sources that collaboratively forms a pool of data. The reason behind is the evolution in the technology like Internet of Things, Distributed Computing or Cloud Computing that are driven by the applications that are more advance in computation as well as in performance [1]. As a result, at present there is such a huge amount of data that need to be properly managed, accessed, secured and updated. Big Data play a responsive role to fulfill all aspects of requirements. The data generated from different sources are big as well as complicated enough that only human efforts are not capable for their inferential analysis. By letting the machines to learn things and think as humans do, such problems can be sorted more accurately and in less time. Hence, "Machine Learning" as the name suggests making a machine that much capable to discover knowledge and give out intelligent decisions from the data sets, comes into play. Based on the learning methodology, a machine can adopt supervised, unsupervised or reinforcement learning [2]. A brief introduction of these three types of learning techniques is discussed in this section. More emphasis is given to unsupervised learning techniques that include data clustering.

1.1 Machine Learning Techniques

Machine Learning is a branch of artificial intelligence [3] that mainly includes the techniques that are capable enough to make a computer system learn from the existing patterns and take decisions [4]. As mentioned above, learning mechanism is divided into three classes: supervised learning, unsupervised learning, and reinforcement learning [2] as shown in the below Figure 1.

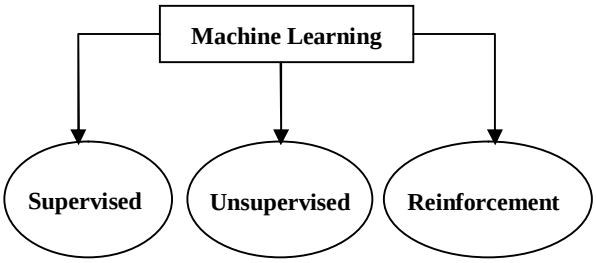

Figure 1: Machine Learning Classification

A. Supervised Learning

Supervised learning, name itself implies that performance of learning algorithm is analyzed by the supervisor to decide whether the decision is good or bad. The dataset is completely labeled i.e., class labels are already known and the learning algorithm verify whether, the performed action is correct or not. Commonly used algorithms that fall in this category are: Support Vector Machine [5], Random Forest [6], Neural Network [7], Regression [8] and Classification [9].

B. Unsupervised Learning

In this category, there is no prior knowledge of the labels in the dataset. The algorithms are designed in such a way that they are capable of finding the similarity between the data items and then, define the labels. Commonly used algorithms in this category are: Clustering Algorithms [10], Self-Organizing Neural Networks [11].

C. Reinforcement Learning

Reinforcement Learning is the very generic learning framework for sequential decisions i.e., models under this category are capable of generating sequence of decisions. In this learning technique, software agents are trained in such a manner that they can take actions in a given environment so that they can maximize the cumulative reward in that particular environment. Some reinforcement learning algorithms are: Q learning, Deep Q Network [12], Deep Deterministic Policy Gradient [13].

The main focus of this paper is towards the unsupervised learning techniques specific to data clustering. Various cluster validation methods are discussed in the next section followed by an experimental analysis to compare the performance of different cluster validity indexes in the further sections.

1.2 Clustering

Clustering is defined as an unsupervised learning technique that groups the data objects without the prior knowledge of the labels or classes [14]. The grouping or clustering of data items is based on the principle that similarity of data items within the cluster or group (intra-cluster similarity) should be maximum and similarity of data items among others clusters (inter-cluster similarity) should be minimum[15]. Among various data analysis techniques, clustering has its wide applications in the field of pattern recognition, information retrieval, image processing, market research etc [16].

Currently, there are numerous clustering algorithms to deal with the different categories of datasets. During the development phase of these algorithms, the processing efficiency, accuracy and performance is enhanced in a very steep manner. The clustering algorithms are divided into five main categories: the partitioning-based algorithms [10], the hierarchical-based algorithms [12], density-based algorithms [17], grid-based algorithms [18] and the model-based algorithms [19]. Among all mentioned clustering algorithms, the partitioning-based algorithms are most prominent. K-means partitioning algorithm in this category has gain popularity because it is simple to implement and is more effective. K-means algorithm lacks stability because of random selection of the initial cluster centers and due to different parameter settings; it may produce different clusters for the same dataset. Later on, several improvements have done to improve the selection of initial cluster centers [20] [21] by predefining the value of K (number of clusters).

1.3 Cluster Validity Index

The empirical rule$(2 \leq K_{max} \leq \sqrt{n})$, gives the number of clusters for dataset that always lies in a fuzzy interval. Hence, practically it is difficult to identify the value of K-optimal (Optimal number of clusters for a dataset.) [22]. Cluster Validity Index (CVI) has been developed to validate cluster as well as to identify the K-optimal value for a given dataset [23]. For every different cluster number, the clustering results are evaluated respectively. Cluster with optimal index value will correspond to the optimal partition of the dataset. The several CVIs used are: Dunn's Index [24] (DI-index), Davies-Bouldin Index [25] (DBI-index), Ibai Gurrutxaga Index [26] (COP-Index), Calinski Harabasz Index [27] (CH-index), Bandyopadhy Index [28] (I-Index), Variance based Clustering Validity Index [29] (VCVI) and BCVI [30].

In the following sections, a brief literature study, methodology used and experimental analysis has been done over two UCI machine learning datasets and two simulated datasets. Results are presented in a tabular form for the sake of simplicity.

2. RELATED STUDY

The partition based algorithms divide the input dataset into different groups that are commonly known as clusters. The groups are made in such a way that the data items within a group are as similar as possible and data items in different groups are dissimilar. In order to evaluate how good a partition, Cluster Validity Indexes are used. CVIs are the key aspect that helps in optimizing and determining the K-Optimal [31]. Numerous CVIs have been proposed in order to make significant evaluation of clusters. For simplicity, CVIs are categorized into three types: CVIs based on the fuzzy division of datasets [32], CVIs based on the statistical knowledge of the dataset [33] and CVIs based on the geometry of the datasets. Xie-Beni [32] CVI is a fuzzy based method that collaborate objective function, structure of the dataset and degree of membership to evaluate the cluster. The basic limitation of fuzzy based CVIs is their poor performance on the results of hard clustering algorithms [34]. In-Group Proportion (IGP) [33] CVI is based on the statistical knowledge of the datasets. IGP uses intra cluster ratio of all the data points to evaluate the cluster performance and hence, it is not suitable for large datasets to determine the K-Optimal [35]. Numerous CVIs have been proposed based on the geometric structure of datasets (DI-index, DBI-index, COP-Index, CH-index, I-index, VCVI, BCVI etc.). Most of the CVIs in this category rely on the assumptions that partition of dataset into clusters is already in the optimal form. However, in most of the cases the optimal clusters are not known [6].

In today's world, dimensionality and size of the dataset is large enough that result in high computational and imbalanced performance of the CVIs [36]. Therefore, it's hard to find the K-Optimal value efficiently for all the datasets. For a spherical distributed dataset, CVIs effectively measures the coherence within the cluster and separation between the clusters [37]. However, for datasets having non-spherical distribution of data items, datasets containing outliers,

datasets containing overlapping values and datasets having variant cluster sizes or densities; it is difficult to find K-Optimal value. Undoubtedly, there are the CVIs for the non spherical and datasets having large degree of overlapping. But, most of the CVIs are inefficient in terms of effectiveness, computation and accuracy [38]. In [29], the proposed VCVI has overcome all these shortcomings by taking comparatively less computational time and high efficiency. The computational time of VCVI has further improved in [39], a new method BCVI has been proposed that is more optimal and efficient in finding the index values for different clusters.

3. METHODOLOGY USED

The experimentation is carried out over the four publically available datasets (4K2, Aggregation, Iris, Hayes Roth) as shown in the Table 1. The dataset 4K2 has 400 data points and 2 dimensions, the Aggregation dataset has 754 data points and 2 dimensions, the Iris dataset has 150 data points and 4 dimensions and the Hayes Roth dataset has 132 data points and 5 dimensions. Two clustering algorithms viz. Improved K-Means [27] and Grid K-Means [37] are used for the purpose of cluster formation for these four datasets. For the performance validation of the clusters formed, seven well known cluster validation indexes (DI-index, DBI-index, I-index, CH-index, COP-index, VCVI-index and BCVI-index) are used. The empirical rule $(2 \leq K_{max} \leq \sqrt{n})$ is used to find out the range of maximum number of clusters for a dataset. The K-Optimal value for every data set is identified using respective CVIs. Table 1 shows the K-Optimal value, Range of K, corresponding to each dataset. The step by step process is as shown in the Figure 2.

Table 1: Details of datasets

Datasets	Samples (N)	K-Optimal	Range of K
4K2	400	4	[2, 20]
Aggregation	754	6	[2, 27]
Iris	150	3	[2, 12]
Hayes Roth	132	3	[2, 11]

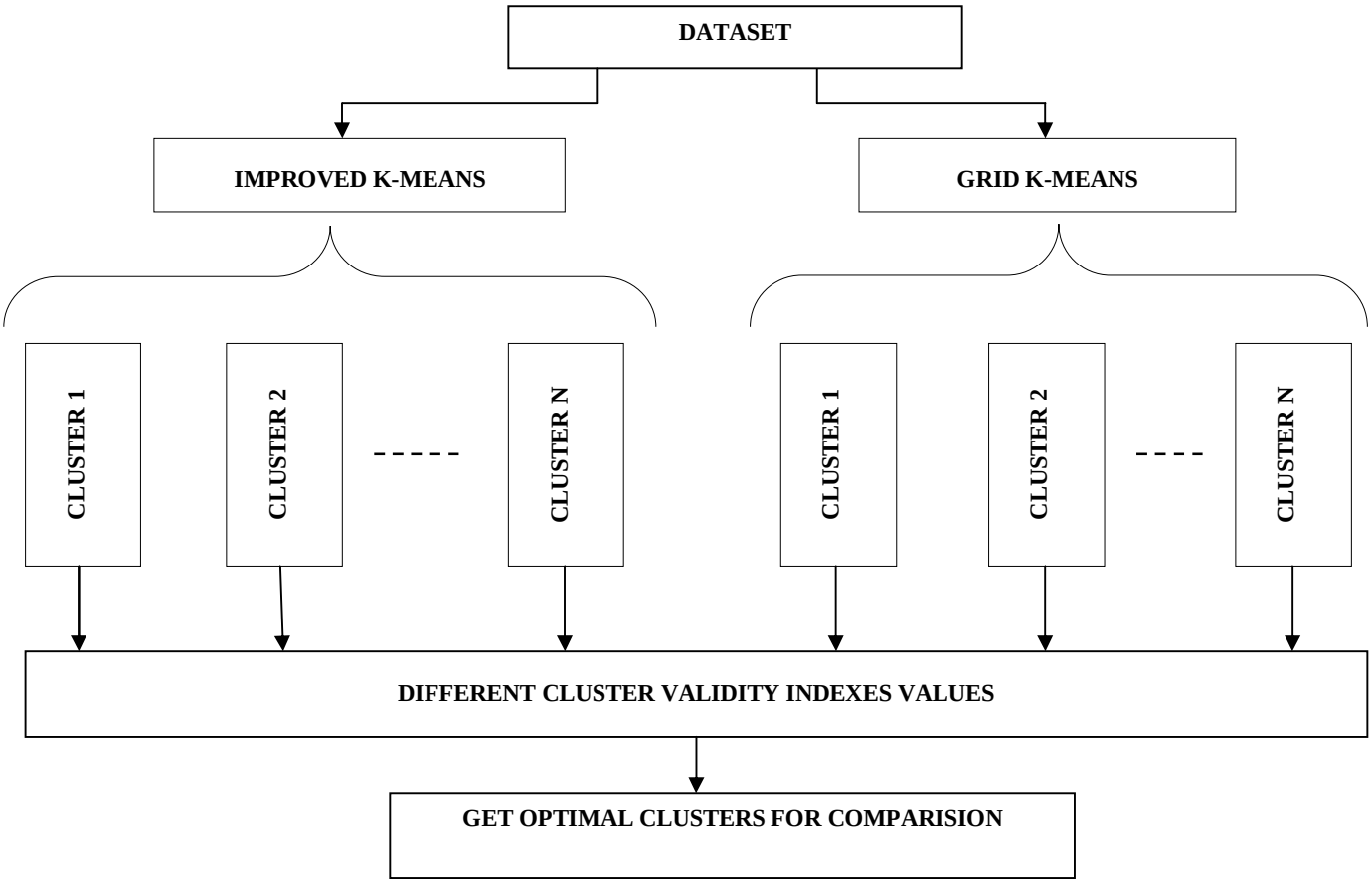

Figure 2: Framework used for analysis

Table 2: Standardized CVI values and K-Optimal values of different datasets

Dataset		4K2		Aggregation		Iris		Hayes-Roth	
		K-Optimal	CVI	K-Optimal	CVI	K-Optimal	CVI	K-Optimal	CVI
Improved K-Means	DI-Index	2	30	24	30	11	30	9	30
	DBI-Index	4	9.6439	4	18.772	2	10.423	2	25.231
	I-Index	3	30	2	30	3	30	2	30
	CH-Index	4	30	18	30	3	30	11	30
	COP-Index	4	15.387	4	23.798	2	14.713	2	25.866
	VCVI-Index	4	1.7574	6	4.422	3	5.687	3	5.721
	BCVI-Index	4	1.6384	6	4.324	3	4.442	3	6.432
Grid K-Means	DI-Index	4	150	25	150	11	150	10	150
	DBI-Index	4	45.242	4	95.266	2	46.635	2	76.80
	I-Index	2	150	4	150	2	150	2	150
	CH-Index	4	150	26	150	2	150	11	150
	COP-Index	4	70.326	5	77.627	2	46.695	10	121.21
	VCVI-Index	4	9.548	6	22.109	3	19.91	3	15.521
	BCVI-Index	4	5.260	6	11.109	3	8.91	3	17.723

4. EXPERIMENTAL ANALSIS

Table 2 shows the experimental results. For a particular dataset, results are obtained for all the clusters within the range of K and only the optimal values are presented in the Table 2. The measured values are used for the comparative analysis of six different CVIs with respect to Improved K-Means and Grid K-Means. For sake of simplicity, values in the Table 2 are graphically shown in the below figures.

4.1 Analysis of 4K2 Dataset

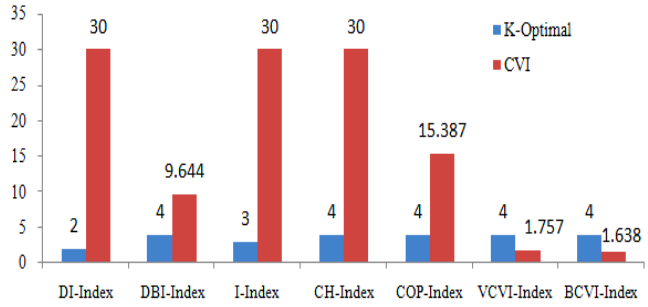

Figure 3: Standardized CVI values of 4K2 dataset (Improved K-Means)

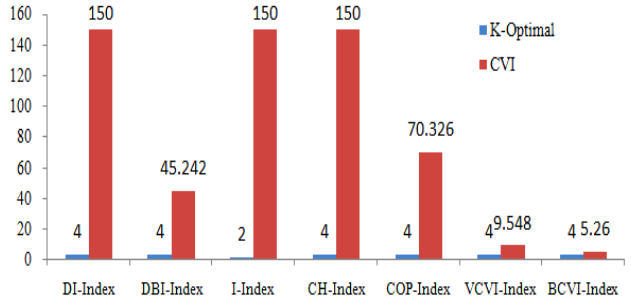

Figure 4: Standardized CVI values of 4K2 dataset (Grid K-Means)

The 4K2 dataset contains 400 sample points for which the range of K lies in an interval of [2, 20] as per the empirical rule. The evaluated CVI values for dataset 4K2 are graphically represented in the Figure 3 and Figure 4. The K-Optimal value is plotted along with CVI values for different indexes using Improved K-Means and Grid K-Means. The DI-Index (2, 30) and I-index (3, 30) are unable to find the optimal clusters for dataset 4K2 when partitioning is done using Improve K-Means as well as (DI-Index (4,150) and I-index (2, 150)) when partitioning is done using Grid K-Means. The optimal clusters for 4K2 dataset obtained from remaining five indexes are plotted in the Figure 3 and Figure 4 by using both Improved K-Means and Grid K-Means respectively. The VCVI and BCVI index shows better performance among all other CVIs. Both VCVI and BCVI have got four optimal clusters with CVI value 1.757 and 1.638 respectively for Improved K-Mean whereas for Grid K-Means, VCVI and BCVI have got 9.548 and 5.26 respectively.

4.2 Analysis of Aggregation Dataset

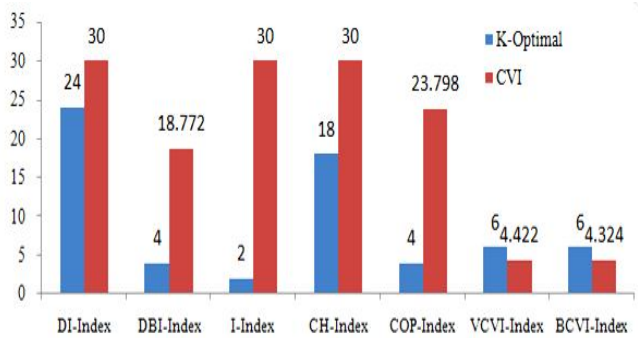

Figure 5: Standardized CVI values of Aggregation dataset (Improved K-Means)

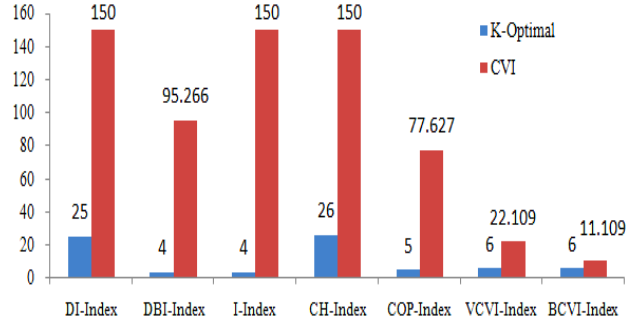

Figure 6: Standardized CVI values of Aggregation dataset (Grid K-Means)

The Aggregation dataset contains the 754 samples and the range of K using empirical rule is [2, 27]. From Figure 5 and Figure 6, it has been observed that the DI-Index, CH-Index and I-Index are unable to get the optimal number of clusters for both Improved K-Means and Grid K-Means. The VCVI has got six optimal clusters for both Improved K-Means and Grid K-Means with CVI values 4.422 and 22.109 respectively. The BCVI has got same six optimal clusters for both Improved K-Means and Grid K-Means with CVI values 4.324 and 11.109 respectively.

4.3 Analysis of Iris Dataset

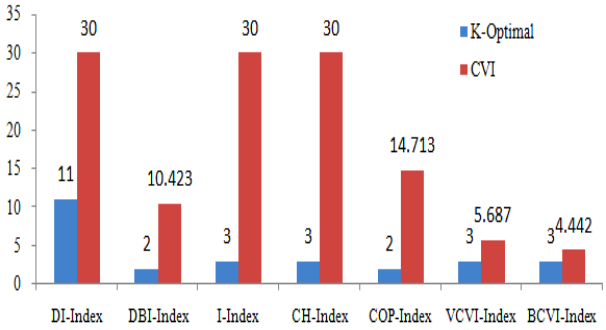

Figure 7: Standardized CVI values of Iris dataset (Improved K-Means)

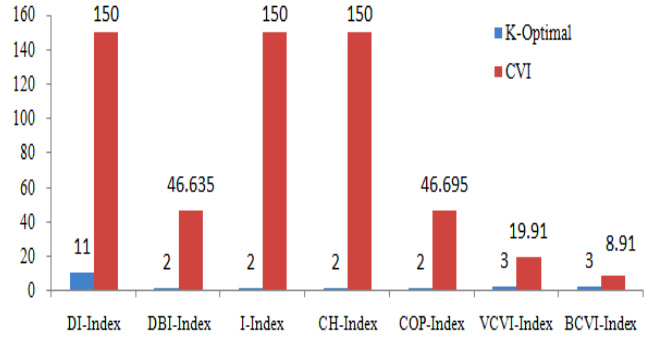

Figure 8: Standardized CVI values of Iris dataset (Grid K-Means)

The Iris dataset contains 150 sample points and the range of K is [2, 12]. Figure 7 depict that CH-Index, I-index, VCVI-Index and BCVI- Index are able to find the optimal clusters but only VCVI and BCVI has got the lowest CVI values for Improved K-Means clustering. The DI-Index is unable to find the optimal cluster number where as the DBI-Index and COP-Index is able to obtain near optimal cluster partition with CVI value 10.423 and 14.713 respectively. Figure 8 shows that only VCVI and BCVI are able to get the optimal clusters with CVI values 19.91 and 8.91 respectively for Grid K-Means. DI-Index is not able to get the optimal cluster number where as DBI-Index, I-Index, CH-Index and COP-Index is able to get the near optimal clustering partition.

4.4 Analysis of Hayes-Roth Dataset

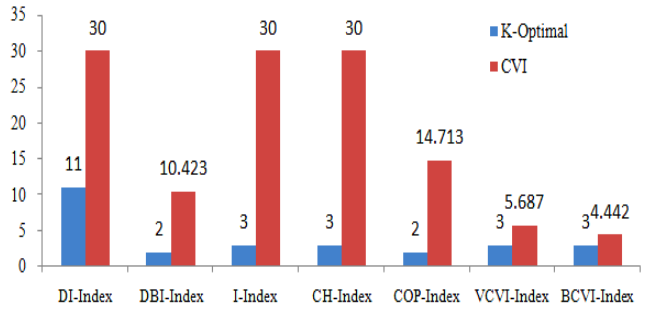

Figure 9: Standardized CVI values of Hayes Roth dataset (Improved K-Means)

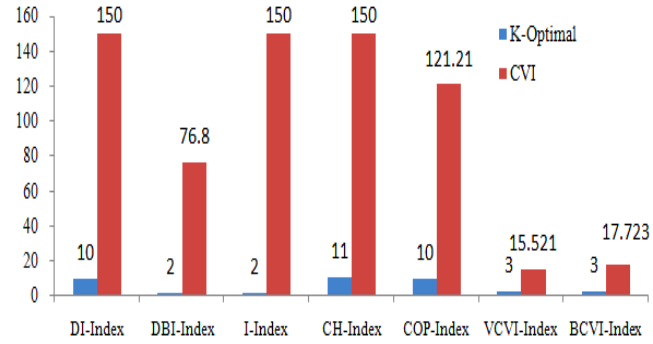

Figure 10: Standardized CVI values of Hayes Roth dataset (Grid K-Means)

The Hayes Roth dataset has 132 sample points with five dimensions. The range of K lies in interval of [2, 11] according to empirical rule. The calculated standard values of different CVIs are listed in Table 2 and are plotted in Figure 9 and Figure 10. In Figure 9, it has been observed that VCVI and BCVI are able to get the optimal clustering partitioning with CVI values 5.721 and 6.432 respectively. The BCVI-Index has got higher CVI value as compared to the VCVI-Index. DBI-Index, I-Index and COP-Index is able to get the near optimal clusters whereas DI-Index and CH-Index are not able to get the optimal cluster partitioning for Improved K-Means. Figure 10 clearly shows that VCVI-Index and BCVI-Index are able to get the optimal clustering partitioning with CVI values 15.521 and 17.723 respectively. Again, the VCVI-Index has shown lower CVI value among all other CVI indexes for Grid K-Means. DBI-Index and I-Index are able to get the near optimal

partition where as COP-Index, CH-Index, DI-Index are not able to get the optimal clustering.

5. CONCLUSION AND FUTURE WORK

The partitioning based clustering algorithms are the most commonly used algorithms in the unsupervised machine learning technique. Irrespective of various clustering algorithms, different CVIs shows different results for the same dataset. The availability of numerous CVIs creates dilemma for a data analyst selection of an appropriate CVI for evaluation. Also, in traditional partitioning algorithms need to set the value of K (number of clusters) in advance. So, an optimal value of K (K-Optimal) for a dataset is very hard to find in traditional cluster partitioning algorithms. Improved K-Means and Grid K-Means algorithms overcome this limitation and help in finding the value of K-Optimal in a very effective manner. In this paper, the performance of seven CVIs is evaluated for four publically available datasets. Moreover, two distinct algorithms viz. Improved K-means and Grid K-Means are used for cluster formation. From the experimental analysis, it has been observed that VCVI-Index and BCVI-Index out performs among all other CVIs. In future, this work can be further extended to analyze large datasets and some heuristic techniques can be used to form clusters.

REFERENCES

1. A. Botta, W. D. Donato, V. Persico, & A. Pescapè, "Integration of Cloud computing and Internet of Things: A survey," *Future Generation Comp. Syst.*, vol. 56, pp. 684-700, 2016. https://doi.org/10.1016/j.future.2015.09.021
2. J. Qiu, Q. Wu, G.Ding, Y. Xu, S.Feng, "A survey of machine learning for big data processing," *Adv. Signal Process*, pp. 1–16, 2016.
3. V. Mani, R. GokulPrasath, S. Jegatha, K. Kannathal, K. MadhuMitha, "Hospital recommendation system using machine learning," *International Journal of Advance Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1324-1327, 2020. https://doi.org/10.30534/ijatcse/2020/64922020
4. S. Mirjalili, H.Faris, I. Aljarah, "Introduction to Evolutionary Machine Learning Techniques." In: Mirjalili S., Faris H., Aljarah I. (eds) *Evolutionary Machine Learning Techniques. Algorithms for Intelligent Systems. Springer, Singapore*, 2020.
5. L. Wang, "Support vector machines: theory and applications," *Springer Science & Business Media*, vol. 177, 2005.
6. P. Probst, M. N. Wright & A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews", *Data Mining and Knowledge Discovery*, vol. 9 no. 3, 2019. https://doi.org/10.1002/widm.1301
7. G.P.Zhang, "Neural networks for classification: a survey". *IEEE Trans Syst Man Cybern Part C (Appl Rev)* vol. 30, pp. 451–462, 2000.
8. I. Gkioulekas, L. G. Papageorgiou, "Piecewise regression analysis through information criteria using mathematical programming," *Expert Systems with Applications*, vol. 121, pp. 362-372, 2019.
9. A. Jake, C. S. Long, B. P. Smith, "Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes." *Expert Systems with Applications*, vol. 115, pp. 245-255, 2019. https://doi.org/10.1016/j.eswa.2018.08.002
10. Z. Kejia, Y. F. Yanan Hu, C. Li, and Panchi Li. "Scheduling strategy for computational-intensive data flow in generalized cluster environments." *Applied Soft Computing*, vol. 82, 2019.
11. S. Maciej, M. Wolkiewicz, T. Orlowska-Kowalska, and C. T. Kowalski. "Application of Self-Organizing Neural Networks to Electrical Fault Classification in Induction Motors." *Applied Sciences*, vol. 9, no. 4, 2019.
12. I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, A. Ignateva, "Deep attention recurrent Q-network." *arXiv preprint arXiv:1512.01693*, 2015.
13. B. Maron, G. Hoffman, MW. Budden D, Dabney W, Horgan D,A. Muldal, "Distributed distributional deterministic policy gradients." *arXiv preprint arXiv:1804.08617*, 2018.
14. J. Huang, Z.L. Yu, Z. Gu, "A clustering method based on extreme learning machine," *Neurocomputing*, vol. 277, pp. 108–119, 2018. https://doi.org/10.1016/j.neucom.2017.02.100
15. S. Mohamed, "Understanding user's behavior by social media data clustering" *International Journal of Advance Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 167-170, 2020. https://doi.org/10.30534/ijatcse/2020/25912020
16. F. Nie, X. Wang, M.I. Jordan, H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering", *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016), Phoenix, USA*, pp. 1969–1976, February 2016.
17. F. Nie, C. Ding, D. Luo, H. Huang, "Improved MinMax cut graph clustering with nonnegative relaxation," *in: Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2010): Part II, Barcelona, Spain*, pp. 451–466, September 2010.
18. R. Xu, D. Wunsch II, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 654–678, 2005.

19. A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", *IEEE Transactions on Emerging Topics in Computing, vol. 2*, pp. 267-279, 2014.
20. M. Erisoglu, N. Calis, S. Sakallioglu, "A new algorithm for initial cluster centers in k-means algorithm," *Pattern Recognit. Lett.*, vol. 32 no. 14, pp. 1701–1705, 2011.
21. Y. Liu, Z. Ma, F. Yu, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy", *Knowl. Based Syst.*, vol. 133, pp. 208–220, 2017.
22. J. Liang, X. Zhao, D. Li, F. Cao, C. Dang, "Determining the number of clusters using information entropy for mixed data," *Pattern Recognit*., vol. 45, no. 6, pp. 2251–2265, 2012. https://doi.org/10.1016/j.patcog.2011.12.017
23. S. Yue, J. Wang, J. Wang, X. Bao, "A new validity index for evaluating the clustering results by partitional clustering algorithms," *Soft Comput.*, vol. 20, no. 3, pp. 1127–1138, 2016.
24. J.C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.* vol. 3, no 3, pp. 32–57, 1973.
25. D.L. Davies, B. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal.Mach. Intell*., vol. 1, no. 2, pp. 224–227, 1979.
26. I. Gurrutxaga, I. Albisua, O. Arbelaitz, J.I. Martin, J. Muguerza, J.M. Perez, I. Perona, SEP/COP: "An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index," *Pattern Recognit.*, vol. 43, no.10, pp. 3364–3373, 2010. https://doi.org/10.1016/j.patcog.2010.04.021
27. T. Caliski, J. Harabasz, "A dendrite method for cluster analysis", *Commun. Stat.*, vol. 3 no.1, pp. 1–27, 1974.
28. U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 12, no. 24, pp. 1650–1654, 2001.
29. Z. Erzhou and Ma, R., "An effective partitional clustering algorithm based on new clustering validity index." *Applied Soft Computing*, vol. 71, pp. 608-621, 2018.
30. Z. Erzhou. and Ma, R. "Effective Clustering Analysis Based on New Designed Clustering Validity Index and Revised K-Means Algorithm for Big Data." *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications*, pp. 96-102, 2018.
31. S. Yue, J. Wang, J. Wang, X. Bao, "A new validity index for evaluating the clustering results by partitional clustering algorithms," *Soft Computing*, vol. 20, no. 3, pp. 1127–1138, 2016. https://doi.org/10.1007/s00500-014-1577-1
32. X.L. Xie, G. Beni, "A validity measure for fuzzy clustering", *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 13, no. 8, pp. 841–847, 1991.
33. A.V. Kapp, "Are clusters found in one dataset present another dataset?", *Biostatistics*, vol. 8, no. 1, pp. 9–31, 2006.
34. F. Nie, D. Xu, X. Li, "Initialization independent clustering with actively self-training method," *IEEE Trans. Syst. Man Cybern. Part B: Cybern*, vol. 42, no. 1, pp. 17–27, 2012.
35. F. Nie, X. Wang, H. Huang, "Clustering and projected clustering with adaptive neighbors", *in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014), New York City, USA*, pp. 977–986, August 2014.
36. O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Perez, I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, 2013.
37. S. Bandyopadhyay, U. Maulk, M.K. Pakhira, "Clustering using simulated annealing with probabilistic redistribution," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 15, no. 2, pp. 269–285, 2001. https://doi.org/10.1142/S0218001401000927
38. R.J.G.B. Campello, Generalized external indexes for comparing data partitions with overlapping categories, *Pattern Recognit. Lett*., vol. 31, no. 9, pp. 966–975, 2010.
39. Z. Erzhou, Y. Zhang, P. Wen, and F. Liu. "Fast and stable clustering analysis based on Grid-mapping K-means algorithm and new clustering validity index," *Neurocomputing*, vol. 363, pp. 149-170, 2019. https://doi.org/10.1016/j.neucom.2019.07.048