VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

Digital Assignment-I, AUGUST 2020

B.Tech., Fall-2020-2021

NAME	PRIYAL BHARDWAJ
REG. NO.	18BIT0272
COURSE CODE	ITE2011
COURSE NAME	MACHINE LEARNING
SLOT	E2+TE2
FACULTY	Prof. DURAI RAJ VINCENT P.M.

CLUSTURING APPROACHES

(All papers taken from IEEE)

Journal 1: Unsupervised K-Means Clustering Algorithm
Authors: Kristina P. Sinaga and Miin-Shen Yang
Date: April 2020
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9072123

In this paper one of the most basic and powerful clustering algorithms in machine learning i.e. K-Means is discussed. In normal K-Means approach, the number of clusters to be formed is known prior to using the algorithm. The authors propose an unsupervised K-Means clustering algorithm that can construct an optimal number of clusters without prior knowledge. X-Means algorithm is used usually when the number of clusters is not known. But even this method requires a range of number of clusters based on clustering validity indices. Their learning procedure can do so without any initialization and parameter selection. Numerous real and numerical datasets were used to test their hypotheses. The computational complexity for the unsupervised K-Means algorithm was found to be *O(ncd).* Through seven different examples the efficiency of usual K-Means, unsupervised K-Means (U-K-Means) approaches was compared. It was concluded that U-K-Means approach had faster computational time and less complexity. The proposed method uses all the data points in the dataset as different clusters to solve the initialization problem, it gets rid of any extra clusters. At the end of the research U-K-Means algorithm ended up being superior to other Means related clustering algorithms.

Journal 2: Efficient Clustering Method Based on Density Peaks with Symmetric Neighborhood Relationship
Authors: Chunrong Wu, Jia Lee, Teijiro Isokawa, Jun Yaoi and Yunni Xia
Date: April 2019
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8695694

In this paper, ways to optimize and better the existing density peaks clustering algorithms are discussed. Density Peaks Clustering (DPC) is a clustering approach that identifies cluster centroids of data by setting up a decision graph. A symmetric neighbourhood relationship would help every single data point to be assigned to a cluster. The idea proposed in this paper is to find the symmetric neighbourhood of all the data points and by employing BFS approach identify both small and big clusters. Reverse kNN has a major influence in this research as it helps identify cluster centroids more easily. When the symmetric neighbourhood graphs (SNGs) were made it was observed that if two initial clusters had a close connection then it was likely that they belonged to the same cluster. They also did experiments on real datasets like Iris and Banknote_A. However, they had to use artificial datasets to show the efficiency of DPC-SNR. DPC-SNR could find all clusters correctly and assign all the data points to their respective clusters with benchmarks data. It was concluded that DCP-SNR successfully identifies cluster centroids over all types data regardless of their dimensionalities or distribution. Thus, the new and efficient DPC-SNR approach has the potential to outperform the basic DPC algorithms.

Journal 3: Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data
Authors: Xiao-Dong Wangi, Rung-Ching Chen, Fei Yan, Zhi-Qiang Zeng and Chao-Qun Hong
Date: March 2019
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8672861

In this paper, K-Means clustering algorithm and an advanced and faster version of it are discussed. K-Means clustering is good for low-dimensional data. But with advancement in technology there is a growth in high-dimensional data and requires Fast Adaptive K-Means (FAKM). Seven datasets were taken into consideration. FAKM is then compared with other clustering approaches like KM, LDAKM, OCMKM, MMCKM, OLSDAKM, TRACK, SRDEKM and DEC. The FAKM was first used to evaluate 4D iris dataset. Two evaluation criteria were used to measure performance metrics. The authors tested hypotheses regarding feature selection, optimization etc. but found that the complexity was high for high-dimensional data. So, they proposed their own methods like imposing an adaptive loss function into the objective function while performing feature selection and simple, iterative algorithms for optimization. Through testing it was observed that FAKM had the best clustering accuracy in most of the cases. An additional adaptive loss function was also imposed to reduce impact of redundant features. Finally, a novel FAKM saved the computation over the usual K-Means when the number of features is more.

Journal 4: An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query
Authors: Shan-Shan Li
Date: February 2020
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8985553

DBSCAN happens to be one of the most famous Density based clustering algorithm as it can identify any shape categories in the dataset along with noise points and outliers. But due to brute force Range-Query the method has high complexity and low efficiency. In this paper, the author proposes an improved version based on neighbour similarity using Cover Tree. Toy datasets as well as five large real datasets were used to test the hypotheses. Cover Tree retrieves neighbour for every data point in parallel and uses triangle inequality to filter out unnecessary distance computations thus speeding up the original DBSCAN approach. Nearest neighbour similarity which reduces redundant distance computations and fast parallel nearest neighbour search which speeds up the process of searching neighbours for each point by Cover Tree, were employed. The results from both the original DBSCAN and the proposed method were compared. The results were consistent meaning that the Cover Tree approach gave accurate results. On comparison with the existing ρ-approximate DBSCAN approach, it was observed that the proposed method is still faster and gives more accurate results. It is also much more suitable and reliable for large datasets. The author plans to incorporate more techniques such as map reduction with DBSCAN to speed up the proposed algorithm even more.

Journal 5: Weighted Cluster Ensemble Based on Partition Relevance Analysis with Reduction Step
Authors: Nejc Ilc
Date: June 2020
Link: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9119391

In this paper the author discusses a way to improve the quality of data partitions so as to not compromise the final clustering solution when several ensemble members are brought together to form a consensus. Partition Relevance Analysis (PRA) uses CVI to evaluate weight of the data points in the dataset before fusion. Here they tested 25 genetic and 15 non-genetic datasets and compared 15 consensus functions. The simple clustering algorithms do not satisfy all the assumptions of a useful clustering. Hence, the author proposes a reduction step of PRA to improve the quality of results. To maximize performance, they tested a method for unsupervised reduction of internal CVI involved in PRA. Then they adapted a consensus of functions to work with weight given by the reduction. Lastly, they did extensive experiments by comparing their proposal with other relevant methods on genetic and non-genetic expression datasets. It was found that PRA performs better with the reduction step as compared to other methods that use PRA without the reduction step. Since this approach is based on unsupervised selection of CVI, it can be used in various cluster validations. Thus, it was concluded that the proposed approach gives measurable performance gains as compared to the original methods. The author plans to work on further advances in this approach to further improve quality and performance of the final clustering solution.

****