Analysis of K-Mean and X-Mean Clustering Algorithms Using Ontology-Based Dataset Filtering

M. Rahmah [1†], Muhammad Ahsan Raza [2††], Z. Fauziah [3†], A. Nor Azhar [4†], Muhammad Fahad Raza [5†††], and Binish Raza [6††††]

[†] Faculty of Computing, College of Computing & Applied Science, Universiti Malaysia Pahang, Pekan, Pahang, West Malaysia

[††] Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan

[†††] Department of Computer Science, Riphah International University, Lahore, Pakistan

[††††] Department of Computer Science, Pakistan Institute of Engineering and Technology, Multan, Pakistan

Summary

In the field of computer science, data mining facilitates the extraction of useful knowledge and patterns from a huge amount of data. Various techniques exist in the data mining domain to explore the links, associations, and patterns from data in data warehouses. Among these techniques, clustering is more prominent in analyzing raw and unlabeled data from a large volume of datasets. The clustering mechanism identifies similar features between data objects and arranges them into clusters. In this paper, we have compared the performance of K-Mean and X-Mean clustering algorithms using two datasets of student enrollment in higher education institutions. Our methodology incorporated ontology to filter the datasets and exploited Rapidminer environment to evaluate the performance of clustering algorithms. The results showed that X-Mean is more suitable for large datasets in terms of discovery and accuracy of clusters.

Key words:

Clustering algorithm, Dataset filtering, Data mining, Ontology construction

1. Introduction

Data mining is the process of discovering information and useful pattern from the large volume of data. The discovery process is iterative and consists of a sequence of steps to identify patterns such as data selection, transformation, cleaning, and patterns, and knowledge evaluation. Through data mining, the hidden relationships between large data are identified which are then summarized in such a way that it helps businesses to make future decisions. Data mining appears to be useful for multiple fields including artificial intelligence, statistics, and database management, to analyze large data stored in data warehouses [1].

Clustering is a common and primary technique in the domain of data mining to facilitate the analysis of stored data. This technique involves the splitting of unlabeled or raw data into groups on the behalf of similarities between data objects. These groups are known as clusters. Authors in [2] suggested that the data objects are made part of a cluster on the basis of two rules: minimizing the inter-class similarity and maximizing the intra-class similarity. These rules simply states that a cluster contains objects which have some sort of similarity between them, while dissimilar objects belong to different clusters. Data clustering which does not require any labeled data (i.e., trained data), the machine makes clusters on the basis of similarities and dissimilarities between the objects. This type of clustering is known as the unsupervised machine learning process. Although supervised data classification is a better approach for dividing data, it requires training data (data is already labeled) which is practically very expensive [3]. In this paper, we aimed to analyze K-Mean and X-Mean algorithms based on clustering efficiency.

The rest of the paper is divided into four sections. The first section highlights the existing research. The second section discusses detail about the proposed methodology. Experimentation and results detail is given in the third section. Finally, the last section presents the conclusion.

2. Literature Review

Singh and Surya [4] presented a comparison of eight clustering algorithms: (i) hierarchal clustering, (ii) K-Means, (iii) M-Tree, (iv) farthest-first traversal, (v) canopy (vi) learning vector quantization, (vii) expectation–maximization (EM), and (viii) density Based Clustering, in term of execution time and no of iterations. The results of the comparison showed that K-mean performed well than the other seven algorithms. Furthermore, density-based algorithm performance is less than the hierarchical algorithm. On the other hand, EM takes more time than other algorithms. Vivek Nailwal et al. [5] presented a review of five algorithms, namely, Grid-based clustering, density-based clustering, K-Mean clustering, hierarchical clustering, and partitioned clustering. The analysis concluded that partitioned clustering works best on the small-size dataset, while density-based clustering achieves better performance on large datasets. Authors in [6]

Manuscript received October 5, 2021
Manuscript revised October 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.10.40

suggested that clustering algorithm performance depends upon the type of domain and data. They compared various clustering techniques including Density-based clustering, Partitioning Clustering, Model-based clustering, hierarchical Clustering, and Grid-based clustering, and results showed that no clustering algorithm works well in all domains. However, K-Mean algorithm performance appeared to be better in many situations.

Saroj and Chaudhary [7] focused on different categories of clustering techniques and reviewed different categories such as Contiguous Clustering (that can be either Nearest Neighbor or Transitive), Well-Separated Clustering, Center-based Clustering, and Density-based Clustering. Fahad et al. [13] presented a survey on clustering algorithms. In this work, the authors provided a comparison of clustering algorithms with two perceptive: theoretical and empirical. Researchers used several validation metrics including, run time and scalability tests to measure the efficiency of clustering algorithms. The overall results showed that empirical clustering algorithms' efficiency is better than others. Rodriguez et al. [14] presented issues related to clustering techniques. For this purpose, the authors used large datasets to address time complexity and accuracy issues regarding clustering algorithms in different environments.

However, no extensive analysis of clustering techniques focused on semantic filtering of datasets (i.e., removal of errors). Because of the strong need for a quality dataset, an analysis of clustering algorithms is need that focus on the semantics of data for the removal of inconsistencies in the dataset. This paper primarily focuses on the construction and use of an ontology to achieve semantic processing of datasets and makes a comparison of clustering algorithms with quality assured datasets.

3. Methodology

To perform a thorough analysis, we have developed a methodology (as depicted in Fig. 1) that consists of three key steps; i) dataset filtering, ii) clustering algorithm, and iii) result and discussion.

3.1 Datasets Filtering

For this research, we have collected two datasets of student enrollment (SE) in institutes from Kaggle [8], which is a community that helps researchers in the data mining field. The chosen datasets are different in size: the first is small and the other is large. The detail about each dataset shown in Table 1.

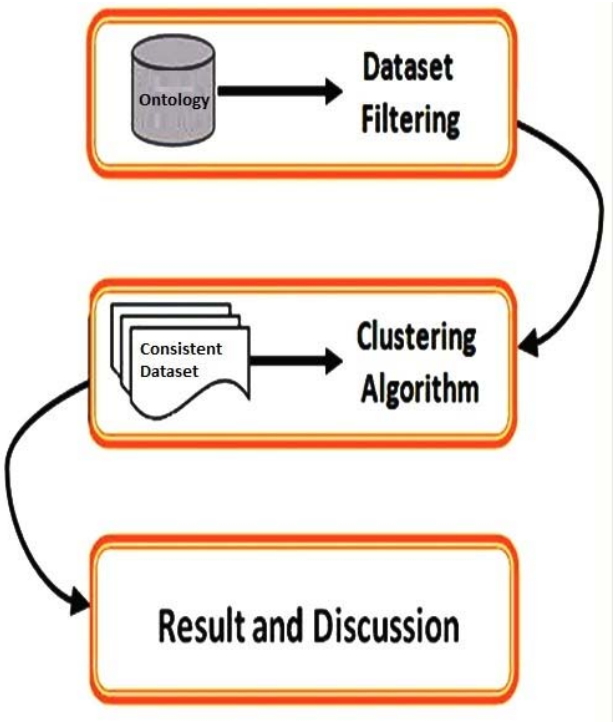

Fig. 1 Steps of methodology

Table 1: Datasets Detail

	Enrollment in HEI	Enrollment in Harvard
No. of Instances	314433	2869
Field	Education	Education

1) SE in Higher Education Institutions (HEI)

This dataset is maintained by Kaggle which stores information is maintained by Kaggle students enrolled in different HEI. This is a larger dataset, whereby all instances (i.e., 314433 rows) and attributes of the dataset are selected to analyze the performance of clustering algorithms (will discuss in section 3.2).

2) SE in Harvard

Harvard university dataset is also taken from the Kaggle community. This is small dataset (it contain 2869 instance) in this research as compared to the first dataset. This dataset is analyzed through Rapidminer environment to measure the performance of chosen clustering algorithms.

3) Ontology-based Data Inconsistencies Removal

In data mining processes, one major issue is the quality of the dataset (i.e., data must be consistent). The existence of inconsistencies in data (for instance, missing values, irrelevant data, and inappropriate data) may lead to useless information in the process of cluster discovery. In literature efforts have been made for the removal of data inconsistencies, thereby improving the quality of data [15, 16]. However, many of these processes inconsistencies manually, which is error-prone and time-consuming. In fact, when dealing with large datasets, such as the HEI dataset (see section 3.1), the manual approach of filtering data inconsistencies becomes unacceptable.

In the context of data quality assurance, researchers have manipulated semantic networks (i.e., ontologies) in the past years for dealing with data inconsistencies. An ontology or semantic network provides a shared description of a domain in a formal way [17, 18], thus facilitate researchers in achieving a consistent state of data. Authors in [20] analyzed the use of ontologies regarding the quality of data and showed improvement in the performance of data mining techniques. Therefore, our methodology also created and utilized ontology for dealing with data inconsistencies. To this end, we have used OntoDataClean system [19] to carry out the cleaning of chosen datasets.

The OntoDataClean system identifies inconsistencies in the dataset on the basis of constructed ontology. It manipulates ontology to obtain information which helps in the detection of errors at the instance-level of the dataset. We have utilized OntoDataClean to filter out three types of data inconsistencies. First, the missing values error is corrected by computing statistical measures such as mean and variance. The MissingValue concept of ontology is used to find divergences among the instances. Second, the typo-errors identification, whereby an instance in the dataset that slightly differs from the correct instance value is detected via InstanceValue concept of the ontology. Third, the outlier errors are identified via InstanceRange concept of the ontology. This concept facilitates in the detection of outlier instances, which are then removed from the dataset.

After the removal of data inconsistencies using ontology, the classification algorithms perform the required and accurate data grouping automatically.

3.2 Clustering Algorithms

In data mining, clustering is a way to divide unlabeled objects into groups on the behalf of their similarities and dissimilarities. Many researchers purposed different clustering algorithms. However, in this research we have selected two popular clustering algorithms: K-Mean and X-Mean, to analyze data of students and evaluate the performance of algorithms.

1) K-Mean

K-Mean is one of the prominent clustering algorithms. The algorithm is a partitioning method that is applied to a dataset to divide data into pre-define distinct non-overlapping groups. The general steps of the algorithm are shown below.

Algorithmic steps for K-Mean clustering [10]
1) Choose the required number of clusters K. 2) Initialization step – to estimate centroids of clusters, choose starting points. 3) Classification step – examines each data object make it part of a cluster base on centroid similarity. 4) Recalculate centroid step – re-calculate new K number of centroids after each data object is examined and allocated to cluster. 5) Repetition – repeat step 3 and step 4 until no change occurs in cluster centroids.

When the K-Mean algorithm applies to a dataset it finds desired number of centroids and makes clusters accordingly to divide data into groups. Partitioning of data into special clusters via K-Mean is done in such a way that each cluster contains objects that remain as close as possible to each other [9]. A typical structure of the K-Mean process is shown in Fig. 2.

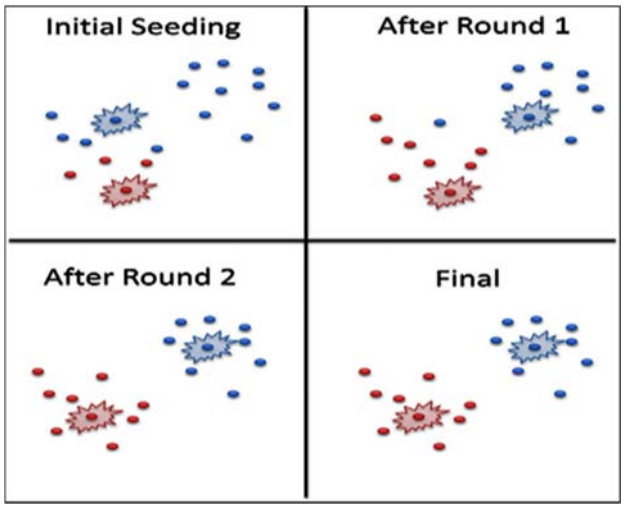

Fig. 2 Typical structure of K-Mean process [11]

2) X-Mean

Pelleg and Moore [12] proposed a new clustering approach know as the X-Mean algorithm. The algorithm is based on the K-mean clustering technique. The proposed algorithm mitigates the issue of the number of clusters in K-mean clustering. Furthermore, it also provides a low computational cost as compared to the K-mean algorithm. X-Mean performs its actions same as done in the first round of K-Mean clustering. Afterward, a local decision has been made to automatically split off the identified centers based on data values. The cluster splitting is done on the basis of BIC criteria.

4. Experimental Results and Discussion

To analyze the performance of clustering techniques, two datasets of different sizes about students' enrollment in different universities are taken from Kaggle (see detail in section 3.1). The Harvard dataset contains a fewer number of instances (2869), whereas the other dataset is large and contains 314433 instances. After removal of inconsistencies from the dataset using an ontology, K-Mean and X-Mean are applied to the datasets via Rapidminer (i.e., an open-source data mining tool) to calculate the efficiency of clustering algorithms. Overall, the experimental results obtained after applying K-Mean and X-Mean clustering algorithms on targeted datasets are shown in Fig. 3 and Fig. 4, respectively. It can be observed from the results that K-Mean shows good performance on the small dataset as compared to the X-Mean algorithm, whereas the X-mean performed better clustering (in terms of size and number of clusters) on a large dataset. Furthermore, it is evident from the results that X-Mean facilitates more amount of student data (opposed to K-Mean clustering) by incorporating an increased number of student groups (i.e., four clusters).

5. Conclusion

Today the usage of a large amount of digit data needs different mining techniques that can extract relevant information from huge data collection. Clustering algorithms play an important role to extract the required information from a large pool of data through different mechanisms. In this paper, two datasets related to student' enrollment in HEI are collected from the Kaggle repository. The first dataset (Harvard data) contains a fewer number of instances as compared to the second one (HEI data). We have incorporated ontology in the process of filtering inconsistencies from the datasets. The use of ontology helps in the removal of data errors in a semantic manner. To classify the instances of consistent datasets (obtained after the semantic filtering of datasets), two clustering algorithms, namely, K-Mean and X-Mean are applied

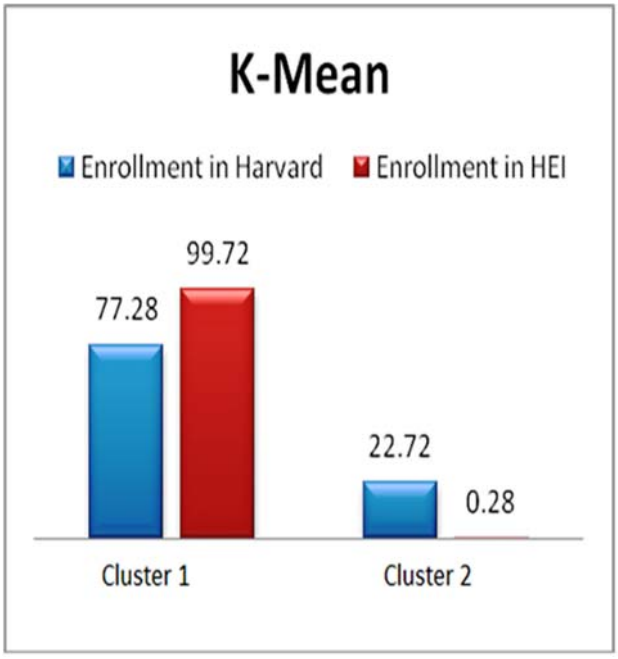

Fig. 3 Performance of K-Mean based on size and number of clusters

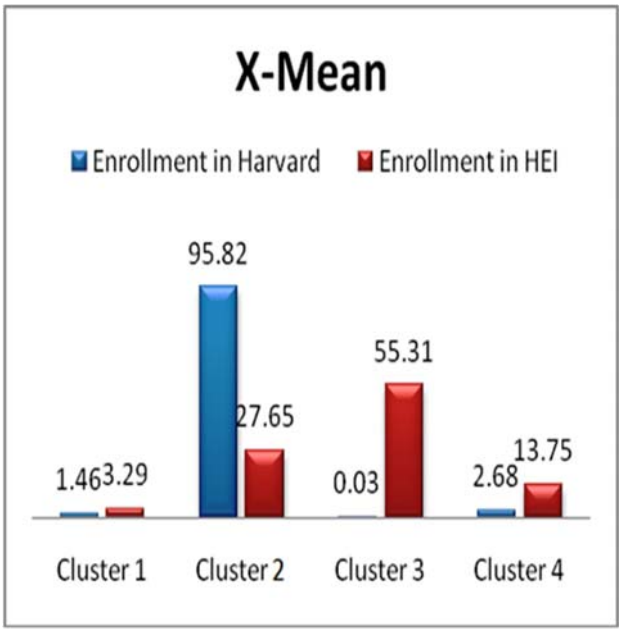

Fig. 4 Performance of X-Mean based on size and number of clusters

through the Rapidminer tool. Later the efficiency of these algorithms in terms of instance classification is computed. After computing efficiency, it is observed that the K-Mean algorithm performed well on the small dataset as compared to the X-Mean algorithm, which depicts better performance of X-Mean over the large datasets. Additionally, it is evident from the results that the X-Mean algorithm (compared to K-Mean clustering) facilitates more amounts of data by incorporating more collection of clusters.

Acknowledgments

The authors would like to thank the Ministry of Higher Education, Malaysia for providing financial support under Fundamental Research Grant Scheme (FRGS) No. FRGS/1/2018/ICT04/UMP/02/1 (University reference RDU190112) and Universiti Malaysia Pahang for research facilities.

References

[1] G. Schuh, G. Reinhart, J.-P. Prote, F. Sauermann, J. Horsthofer, F. Oppolzer, and D. Knoll, "Data mining definitions and applications for the management of production complexity," *Procedia CIRP*, vol. 81, pp. 874-879, 2019.

[2] A. Ahmad, and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," Ieee Access, vol. 7, pp. 31883-31902, 2019.

[3] Aastha Joshi and Rajneet Kaur .: "A Review: Comparative Study of Various Clustering Techniques in Data Mining" IJARCSSE, 2013

[4] Prakash Singh and Aarohi Surya2. "Performance Analysis of Clustering Algorithms in Data Mining in WEKA" IJAET, 2015

[5] Garima, H. Gulati, and P. K. Singh, "Clustering techniques in data mining: A comparison," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 410-415.

[6] D. Patel, R. Modi, and K. Sarvakar, "A Comparative Study of Clustering Data Mining: Techniques and Research Challenges," Int. J. Latest Technol. Eng. Manag. Appl. Sci., vol. 3, no. 9, pp. 67–70, 2014.

[7] Saroj, and T. Chaudhary, "Study on Various Clustering Techniques," International Journal of Computer Science and Information Technologies, vol. 6, no. 3, 2015.

[8] Kaggle. "Datasets," 01 February 2021; https://www.kaggle.com/datasets

[9] Mustakim, U. Khairunnisa, A. Wenda, A. Ilham, F. E. Laumal, A. D. Gs, D. S. Putra, I. B. A. I. Iswara, S. R. Fitriatien, and R. Rahim, "Unsupervised Learning As A Data Sharing Model In The Fp-Growth Algorithm In Determining The Best Transaction Data Pattern," Journal of Theoretical and Applied Information Technology, vol. 99, no. 11, pp. 2679-2689, 2021.

[10] B. Al-Attar, A. J. Allami, A. T. A. Imeer, Y. F. Alasadi, N. M. Norwawi, And H. M. Kadhim, "A Hybrid Approach For Web Search Result Clustering Based On Genetic Algorithm With K-Means," Journal of Theoretical and Applied Information Technology, vol. 99, no. 11, pp. 2722-2733, 2021.

[11] Page, J.T., Liechty, Z.S., Huynh, M.D. *et al.* BamBam: genome sequence analysis tools for biologists. *BMC Res Notes* 7**,** 829 (2014)

[12] D. Pelleg and A. Moore, "X-means: Extending k-means with e_cient estimation of the number of clusters," In: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 727-734.

[13] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, A. Zomaya, I. Khalil, S. Foufou, A. Bouras , "A survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," IEEE 2014, pp. 267 – 279.

[14] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. d. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," PloS one, vol. 14, no. 1, pp. e0210236, 2019.

[15] S. A. Alasadi, and W. S. Bhaya, "Review of data preprocessing techniques in data mining," Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102-4107, 2017.

[16] A. Fatima, N. Nazir, and M. G. Khan, "Data cleaning in data warehouse: A survey of data pre-processing techniques and tools," IJ Information Technology and Computer Science, vol. 3, pp. 50-61, 2017.

[17] M. A. Raza, M. Rahmah, S. Raza, A. Noraziah, and R. A. Hamid, "A Methodology for Engineering Domain Ontology using Entity Relationship Model," *International Journal of Advanced Computer Science and Applications(IJACSA),* vol. 10, no. 8, 2019.

[18] M. A. Hamza, M. J. Ab Aziz, and N. Omar, "Sentence Similarity Measurement for Smart Education Based on Thematic Role and Semantic Network Techniques", ijsecs, vol. 5, no. 2, pp. 37–65, Jan. 2020.

[19] D. Perez-Rey, A. Anguita, and J. Crespo, "OntoDataClean: Ontology-based Integration and Preprocessing of Distributed Data", Lec. notes in Computer Science 4345, 2006, pp. 262-272.

[20] H. Cespivova, J. Rauch, V. Svatek, M. Kejkula, and M. Tomeckova, "Roles of Medical Ontology in Association Mining CRISP-DM Cycle", In Proceedings of the ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa, 2004.