A Study on Some Clustering Algorithms in Data Mining

ICTU Student 171
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student171@ictu.edu.vn

***Abstract*—The thesis surveys data mining and several clustering methods, with particular attention to K-means, CURE, Seeded K-means, and Constrained K-means. Its contents list a comparison of K-means and CURE through three experiments and an application chapter with implementation steps for the constrained method. The available source excerpt documents the structure of these analyses and the author's concluding account of the algorithms studied. It does not provide the experimental datasets, measured scores, or comparative findings needed to establish that one algorithm outperformed another. This summary therefore reports the methods covered and the documented scope without inventing performance results.**

***Index Terms*—Data Mining, Clustering, K-means, Cure, Seeded-KMeans, Constrained-KMeans**

I. Introduction

The thesis begins with an overview of data mining, including its definition, applications, and challenges. It discusses the importance of clustering algorithms in data mining and their significance in both scientific and practical contexts. The author highlights the growing interest in data mining and clustering algorithms, both in Vietnam and globally, and their relevance in various fields. The introduction also outlines the structure of the thesis, which includes a detailed exploration of clustering algorithms and their implementation [1]. Published articles on the methods and subject of this project provide context for this English summary [10].

II. Methods

The author first introduces knowledge discovery and the role of clustering in data mining. The thesis then presents K-means and CURE and identifies an experiment section with three comparisons of those methods. A later application chapter describes problem formulation, implementation, seed-based center initialization, distance functions, and Constrained K-means. Seeded K-means is also included in the survey. The available source excerpt does not expose the details of the three experimental datasets or a reproducible evaluation protocol, so the methods can be summarized at the level of the documented chapter plan and algorithmic topics [1].

III. Results

The documented output is a structured treatment of data mining and the selected clustering algorithms. The source table of contents identifies three comparisons involving K-means and CURE and an application chapter that implements components of Constrained K-means. In the conclusion, the author states that K-means, CURE, Seeded K-means, and Constrained K-means were presented, while acknowledging limits of time and depth. The available excerpt supplies no dataset size, accuracy score, runtime comparison, or supported ranking among the methods. No numerical or comparative result is inferred from a chapter heading alone [1].

Supervisor: ICTU Teacher 111, Thai Nguyen University of Information and Communication Technology.

IV. Discussion

K-means, CURE, and constrained or seeded variants address different ways of grouping data, but their relative merits cannot be judged from the excerpted conclusion and contents alone. The thesis describes clustering as an important approach to knowledge discovery and identifies several algorithms for study. Its author explicitly notes that time and experience limited the depth of the work. A fuller evaluation would require the experimental data, the measures used, and results for each comparison. The present summary is consequently limited to the documented topics, implementation components, and limitations of the manuscript [1].

V. Conclusion

The thesis concludes that data mining and clustering algorithms are increasingly important in both scientific research and practical applications. The author acknowledges the limitations of the study due to time constraints and expresses a desire for further research in the field. The work highlights the growing interest in data mining and clustering algorithms, both in Vietnam and globally, and their significance in various fields. The author also notes the need for more comprehensive evaluations of clustering algorithms in future studies [1].

References

[1] ICTU Student 171, "A Study on Some Clustering Algorithms in Data Mining," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2010.
[2] Chuyên đề Khai phá dữ liệu và Nhà kho dữ liệu (2004) – TS Đỗ Phúc - Đại học Quốc Gia Thành Phố Hồ Chí Minh
[3] Data Mining: Concepts, Models, Methods, and Algorithms by Mehmed Kantardzic John Wiley & Sons 2003
[4] Basu, S., Banerjee, A., & Mooney, R. J. (2002). Semi-supervised clustering by seeding. In Proceedings of 19th International Conference on Machine Learning (ICML-2002), pp. 19–26

[5] Basu, S., Banerjee, A., & Mooney, R. J. (2003). Active semi-supervision for pairwise constrained clustering. Submitted for publication, available at http://www.cs.utexas.edu/~sugato/

[6] Basu Sugato (2004) Semi-supervised Clustering with Limited Background Knowledge. Proceedings of the Ninth AAAI/SIGART Doctoral Consortium, pp. 979-980, San Jose, CA, July 2004

[7] Blake, C. L., & Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html

[8] Cohn, D., Caruana, R., & McCallum, A. (2000). Semi-supervised clustering with user feedback. Unpublished manuscript. Available at http://www2.cs.cmu.edu/~mccallum/

[9] Kantardzic Mehmed, 2003 DATA MINING:Concepts, Models, Methods & Algorithms. Copyright © 2003 by Mehmed Kantardzic. pp 79-120

[10] J. Kogan, C. Nicholas, V. Volkovich, ”Data Mining - Text mining with information-theoretic clustering,” Computing in Science & Engineering, vol. 5, no. 6, pp. 52-59, 2003. DOI: 10.1109/mcise.2003.1238704.