978-1-5090-3525-0/16/$31.00 ©2016 IEEE

Accelerating Clustering Algorithms Using GPUs

Mahmoud Al-Ayyoub, Qussai Yaseen, Moahmmed Shehab, Yaser Jararweh and Firas Albalas
Jordan University of Science and Technology, Jordan

***Abstract*—Big data is a main problem for data mining methods. Fortunately, the rapid advances in affordable high performance computing platforms such as the Graphics Processing Unit (GPU) have helped researchers in reducing the execution time of many algorithms including data mining algorithms. This paper discusses the utilization of the parallelism capabilities of the GPU to improve the the performance of two common clustering algorithms, which are K-Means (KM) and Fuzzy C-Means (FCM) algorithms.**

I. Introduction

K-Means (KM) and Fuzzy C-Means (FCM), two very common methods for clustering data, face serious issues when they deal with big data [1], [2]. The execution times for these techniques increase as the data size increases, which makes big data clustering a major issue. Furthermore, the number of dimensions may reduce the speed of finishing the clustering operation [3].

To increase the efficiency of clustering algorithms on big data, parallel programming is used. To this end, Graphics Processing Unit (GPUs) are gaining more popularity for compute-intensive computation compared with the Central Processing Units (CPUs). The reason for this is very simple. While modern CPUs can run up to 32 threads at same time, modern GPUs can run around 4999 threads [4]. Obviously, GPUs have higher capabilities to run more threads than CPUs. Therefore, many researchers utilize this advantage to improve the performance of many algorithms [5]. This paper leverages the capabilities of GPUs and parallel techniques in big data clustering. This reduces the effect of increasing data size and number of dimensions, and increases the scalability of applying KM and FCM clustering algorithms.This paper aims to perform fair comparisons using modern CPUs with modern GPUs.

II. Experiments and Results

The specifications of the hardware and software used are as follows.

- Hardware: a 2.20 GHz CPU Intel I7 fourth generation with 6GB RAM. The GPU is NVIDIA GT 740M with 2GB memory.
- Software: 64-bit Windows 10 operating system, CUDA 7.5 toolkit, CUDA drivers and Microsoft visual studio 2013.

Figures 1 and 2 show the effects of increasing the dataset sizes and dimensions on the performance of the two algorithms.

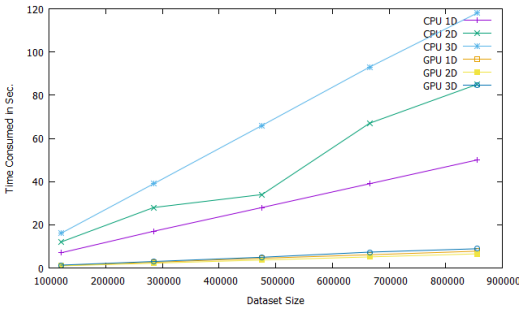

Fig. 1. The performance of different implementations of KM with different dataset sizes and dimensions

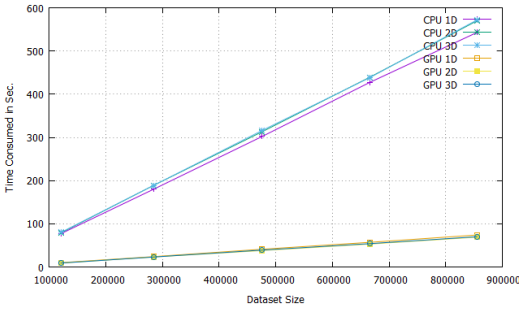

Fig. 2. The performance of different implementations of FCM with different dataset sizes and dimensions

III. Conclusion

This work has presented our effort to study the performance improvements gained by utilizing GPUs to speed up the performance of the two common clustering algorithms, KM and FCM. The experiments aimed to study the effect of increasing data size and number of dimensions on the performance gain of the implementations of the two algorithms in comparing to the sequential ones.

References

[1] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.
[2] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.
[3] S. Ghosh and S. K. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, pp. 34–39, 2013.
[4] S. Cook, *CUDA programming: a developer's guide to parallel computing with GPUs*. Newnes, 2012.
[5] M. A. Shehab, M. Al-Ayyoub, and Y. Jararweh, "Improving fcm and t2fcm algorithms performance using gpus for medical images segmentation," in *Information and Communication Systems (ICICS), 2015 6th International Conference on*. IEEE, 2015, pp. 130–135.