

School of Science
Department of Computer Science and Engineering
Master's Degree in Computer Science

Clustering aggregation on a neutral atom quantum computer

Supervisor:
Prof. Stefano Lodi

Submitted by:
Riccardo Scotti

Cosupervisors:
Dr. Gabriella Bettonte
Dr. Antonio Costantini

Session II
Academic Year 2023/2024

UNIVERSITY OF BOLOGNA

Abstract

School of Science
Department of Computer Science and Engineering

Master's Degree in Computer Science

Clustering aggregation on a neutral atom quantum computer

by Riccardo Scotti

TODO

Acknowledgements

TODO

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
2 Background	3
2.1 Clustering	3
2.1.1 Formal definition	3
2.1.2 Clustering aggregation	3
2.2 Quantum computing	3
2.2.1 Elements of quantum computing	4
2.3 Quantum annealing	4
2.4 Neutral atoms technology	4
3 State of the art	5
4 Proposed method	7
5 Experimental setup	9
5.1 Quantum platforms	9
5.1.1 Pulser	9
5.2 Dataset	9
5.3 Evaluation metrics	9
5.3.1 Silhouette score	9
5.3.2 Rand index	10
5.4 Description of experiments	11
5.4.1 Pulser experiment	11
6 Results	13
6.1 Fresnel experiment	13
7 Conclusions	15
Bibliography	21

Chapter 1

Introduction

TODO

Chapter 2

Background

2.1 Clustering

Clustering is an unsupervised data analysis technique that can be informally defined as the partitioning of a set of objects into groups called clusters; such partitioning must ensure that objects within a same cluster are more similar¹ to each other than to objects in other clusters. [6]

Clustering is widely employed in numerous scenarios that require to group together similar data points, or to extract knowledge from a set of objects in the absence of any prior information. For instance, it is used in marketing and finance as a profiling tool; in image processing and computer vision, as a segmentation technique [5], where it plays a pivotal role in various fields, such as remote image sensing [12] and digital forensics [9]; by energy distribution companies, to optimize the allocation of resources to end users.

2.1.1 Formal definition

Let $D = \{x_1, \dots, x_n\}$ be a set of objects, also called *dataset*; a clustering for D is a collection of k elements $\mathcal{C} = \{c_1, \dots, c_k\}$, with $\mathcal{C} \subseteq \mathcal{P}(D)$, such that the following properties hold:

$$D = \bigcup_{i=1}^k c_i \quad (2.1)$$

$$c_i \cap c_j = \emptyset \quad \forall c_i, c_j \in \mathcal{C}. \quad (2.2)$$

2.1.2 Clustering aggregation

2.2 Quantum computing

Quantum computing is a set of computational models and paradigms that combine concepts of computer science, physics and engineering. It offers an theoretical framework alternative to that of classical computing, which proved to be beneficial in certain areas of interest, such as cryptography and optimization.

The basic ideas of quantum computing were established by Richard Feynman in a 1982 paper, discussing the usage of computers to perform physics simulations [3]. Feynman observed that, due to their inherent complexity, quantum

¹Provided that a binary ordering relation is defined on the set of objects.

systems could not be feasibly simulated by classical computers, and introduced the idea of computing machines based on quantum principles [2]. A subsequent work by Deutsch demonstrated, via the construction of a quantum Turing machine, that quantum computers are computationally equivalent to classical computers, and could therefore be useful for scopes beyond mere quantum simulation [1].

2.2.1 Elements of quantum computing

The fundamental unit of information in quantum computing is the quantum bit, or qubit for short. Regardless of their physical realization, qubits are characterized by their state, which carries information. While the state of a classical bit is always only one of two logical values (usually called 0 and 1), the state of a qubit works in a significantly different manner, reflecting the underlying principles of quantum mechanics.

Similarly to its classical counterpart, when the state of a qubit $|\psi\rangle$ is measured, it can assume either one of two logical values (usually called $|0\rangle$ and $|1\rangle$); the key difference with a bit is that the state of a qubit before it is measured is a linear combination of the two possible states:

$$|\psi\rangle = \alpha |0\rangle + \beta |1\rangle, \quad (2.3)$$

where (α, β) is a unit vector of a Hilbert space or, in other words, $|\alpha|^2 + |\beta|^2 = 1$.

2.3 Quantum annealing

2.4 Neutral atoms technology

Chapter 3

State of the art

Chapter 4

Proposed method

Chapter 5

Experimental setup

This chapter provides a description of the experimental setup used to evaluate the clustering aggregation protocol in different quantum computing environments, both on a simulator and on real hardware. The protocol was tested on three distinct platforms: a simulator for a neutral atom quantum computer, specifically the Pulser simulator by PASQAL; the Fresnel neutral atom quantum computer, also developed by PASQAL; the Advantage quantum annealer with Pegasus topology, developed by D-Wave.

5.1 Quantum platforms

5.1.1 Pulser

5.2 Dataset

The dataset used to test the protocol is shown in figure 5.1. It was first introduced by Gionis, Mannila and Tsaparas to test classical clustering aggregation algorithms [4]; it was then used by Li and Latecki to test a clustering aggregation protocol that uses simulated annealing [8].

It is specifically designed to contain different shapes and configuration of points, in such a way that most clustering algorithms fail to produce a correct clustering. It is made up of 7 clusters and 788 points in total.

5.3 Evaluation metrics

Different metrics were used to evaluate the quality of single clusters and of clustering algorithms as a whole. Silhouette score was used as a factor to weigh clusters in the clustering aggregation protocol; Rand index was used to compare the overall quality of the clustering algorithms and aggregation protocols.

5.3.1 Silhouette score

The silhouette score is a metric used to evaluate clustering quality, introduced by Rousseeuw [11]. The score is computed for each point in the dataset and reflects both its cohesion with points in the same cluster, as well as the separation with

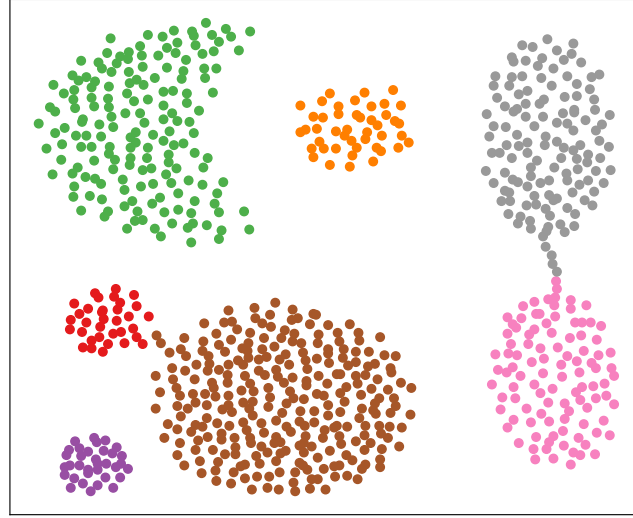


Figure 5.1: Dataset plot

respect to points in different clusters. Specifically, the silhouette score of a point p in the dataset is defined as

$$S_p = \frac{b_p - a_p}{\max(a_p, b_p)}, \quad (5.1)$$

where a_i is the average distance from the point to other points in the same cluster (intra-cluster distance), and b_i is the average distance to points in the nearest different cluster (inter-cluster distance). The score ranges between -1 and 1, where values close to 1 indicate that points are well-clustered, with high cohesion and good separation, while negative values suggest points may have been assigned to the wrong cluster.

A quantification of the quality for a single cluster can be obtained by computing the average of the silhouette score of the point it contains; more formally, for a cluster c_i , the formula for its average silhouette (AS_i) is

$$AS_i = \frac{\sum_{p \in c_i} S_p}{|c_i|}. \quad (5.2)$$

5.3.2 Rand index

The Rand index is a metric that quantifies the similarity between two data clusterings, first proposed by Rand [10]. It is computed by considering all possible pairs of points in the dataset and assessing whether they are assigned to the same or different clusters in the two clusterings being compared. The index assumes values between 0 and 1, with 1 indicating perfect agreement between the two clusterings.

Given two clusterings \mathcal{C}_1 and \mathcal{C}_2 , their Rand index is computed as

$$RI = \frac{a + b}{a + b + c + d}, \quad (5.3)$$

where

- a is the number of point pairs that are put in the same cluster in both clusterings;
- b is the number of point pairs that are put in different clusters in both clusterings;
- c is the number of point pairs that are put in the same cluster in \mathcal{C}_1 , but in different clusters in \mathcal{C}_2 ;
- d is the number of point pairs that are put in different clusters in \mathcal{C}_1 , but in the same cluster in \mathcal{C}_2 .

5.4 Description of experiments

5.4.1 Pulser experiment

The Pulser neutral atom computer simulator was used to test the aggregation protocol on the dataset discussed in 5.2;

Two different clustering algorithms were run on the dataset, DBSCAN and Spectral Clustering. Hyperparameters were tuned empirically, in order to ensure an overall amount of clusters inferior or equal to 14, so as not to exceed the amount of qubits the simulator can handle [7].

Chapter 6

Results

6.1 Fresnel experiment

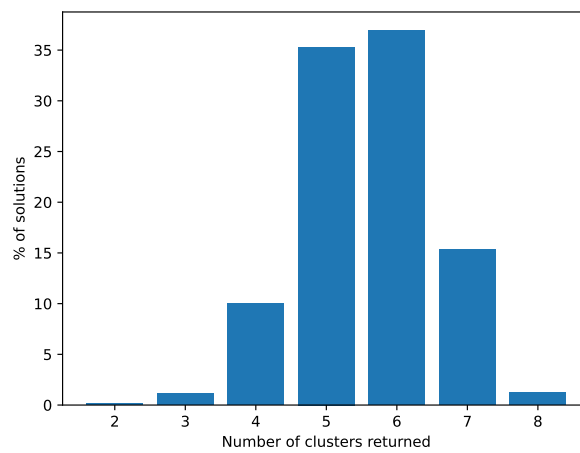


Figure 6.1: Histogram for probability of number of clusters returned in Fresnel experiment

Number of clusters	Probability (%)
2	0.12
3	1.11
4	10.0
5	35.31
6	36.91
7	15.31
8	1.23

Table 6.1: Probability of number of clusters returned in Fresnel experiment

Chapter 7

Conclusions

List of Figures

5.1	Dataset plot	10
6.1	Histogram for probability of number of clusters returned in Fresnel experiment	13

List of Tables

6.1	Probability of number of clusters returned in Fresnel experiment	13
-----	--	----

Bibliography

- [1] David Deutsch. "Quantum theory, the Church–Turing principle and the universal quantum computer". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 400.1818 (1985), pp. 97–117.
- [2] Richard P. Feynman. "Quantum Mechanical Computers". In: *Between Quantum and Cosmos*. Princeton University Press, Dec. 2017, 523–548. DOI: 10 . 1515/9781400886975-036. URL: <http://dx.doi.org/10.1515/9781400886975-036>.
- [3] Richard P. Feynman. "Simulating physics with computers". In: *International Journal of Theoretical Physics* 21.6–7 (June 1982), 467–488. ISSN: 1572-9575. DOI: 10 . 1007/bf02650179. URL: <http://dx.doi.org/10.1007/BF02650179>.
- [4] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. "Clustering aggregation". In: *ACM Trans. Knowl. Discov. Data* 1.1 (2007), 4–es. ISSN: 1556-4681. DOI: 10 . 1145/1217299 . 1217303. URL: <https://doi.org/10.1145/1217299.1217303>.
- [5] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing, Global Edition*. en. 4th ed. London, England: Pearson Education, Sept. 2017, pp. 770–772.
- [6] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review". In: *ACM Comput. Surv.* 31.3 (1999), 264–323. ISSN: 0360-0300. DOI: 10 . 1145 / 331499.331504. URL: <https://doi.org/10.1145/331499.331504>.
- [7] J.R. Johansson, P.D. Nation, and Franco Nori. "QuTiP: An open-source Python framework for the dynamics of open quantum systems". In: *Computer Physics Communications* 183.8 (Aug. 2012), 1760–1772. ISSN: 0010-4655. DOI: 10 . 1016/j . cpc . 2012 . 02 . 021. URL: <http://dx.doi.org/10.1016/j.cpc.2012.02.021>.
- [8] Nan Li and Longin Jan Latecki. "Clustering aggregation as Maximum-Weight Independent Set". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. NIPS'12*. Lake Tahoe, Nevada: Curran Associates Inc., 2012, 782–790.
- [9] Francesco Marra et al. "Blind PRNU-Based Image Clustering for Source Identification". In: *IEEE Transactions on Information Forensics and Security* 12.9 (2017), pp. 2197–2211. DOI: 10 . 1109/TIFS.2017.2701335.
- [10] William M. Rand. "Objective criteria for the evaluation of clustering methods". In: *Journal of the American Statistical Association* 66.336 (1971), pp. 846–850.

-
- [11] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.
 - [12] Anne Solberg, Torfinn Taxt, and Anil Jain. "A Markov Random Field Model for Classification of Multisource Satellite Imagery". In: *Geoscience and Remote Sensing, IEEE Transactions on* 34 (Feb. 1996), pp. 100 –113. DOI: 10 . 1109/36.481897.