

# Balanced K-Means Clustering on an Adiabatic Quantum Computer

Applied Quantum Machine Learning Project



POLITECNICO DI MILANO

June 6, 2021

*Authors:*

Pierriccardo OLIVIERI

Francesco PIRO

Matteo SACCO

*Professors:*

Cremonesi PAOLO

Alessandro LUONGO

Maurizio FERRARI DACREMA

## Introduction

Balanced  $k$ -Mean

Unconstrained  $k$ -Mean Clustering

Balanced  $k$ -means clustering

QUBO formulation

## Analysis

Theoretical

Empirical

Benchmark

## Conclusions

## Critical View

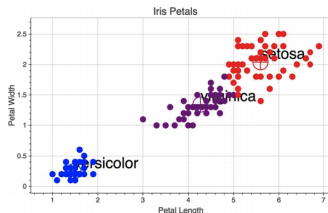


## Outline

### Advantages over classical approaches

- Better targets the global solution of the training problem
- Better theoretic scalability on large datasets

- QUBO formulation and theoretical analysis
- Empirical Analysis
- Authors' Conclusions
- Our Conclusions, opinions and considerations



## Lloyd's algorithm

- Complexity  $O(Nkdi)$  [13]
  - $N$  number of data points
  - $k$  number of clusters
  - $d$  number of features
  - $i$  number of iterations before the algorithm converges

## Scikit-learn implementation

- Complexity  $O(Nkd)$  [18]

[13] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means clustering algorithm" Applied Statistics  
[18] "Scikit-learn: Machine learning in python," J. Mach. Learn. Res.



## Malinen et al.

- Complexity  $O(N^3)$  [13]

---

**Algorithm 1.** Balanced  $k$ -means

Input: dataset  $X$ , number of clusters  $k$

Output: partitioning of dataset.

---

Initialize centroid locations  $C^0$ .

$t \leftarrow 0$

**repeat**

Assignment step:

Calculate edge weights.

Solve an Assignment problem.

Update step:

Calculate new centroid locations  $C^{t+1}$

$t \leftarrow t + 1$

**until** centroid locations do not change.

Output partitioning.

---

[21] Malinen, Mikko. (2014). Balanced K-Means for Clustering.



$$\min_{z \in \mathbb{B}^M} z^T A z$$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$





$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x, y \in \phi_j} \|x - y\|^2$$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x, y \in \phi_j} \|x - y\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x, y \in \phi_j} \|x - y\|^2$$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x, y \in \phi_j} \|x - y\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x, y \in \phi_j} \|x - y\|^2$$

Distance matrix:  $D$

Assignment matrix:  $\hat{W}$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x, y \in \phi_j} \|x - y\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x, y \in \phi_j} \|x - y\|^2$$

Distance matrix:  $D$

Assignment matrix:  $\hat{W}$

$$\sum_{x, y \in \phi_j} \|x - y\|^2 = \hat{w}_j'^T D \hat{w}_j'$$



$$\min_{z \in \mathbb{B}^M} z^T A z$$

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\Phi = \{\phi_1, \phi_2, \dots, \phi_k\}.$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x \in \phi_j} \|x - \mu_j\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x, y \in \phi_j} \|x - y\|^2$$

$$\min_{\Phi} \sum_{j=1}^k \sum_{x, y \in \phi_j} \|x - y\|^2$$

Distance matrix:  $D$

Assignment matrix:  $\hat{W}$

$$\sum_{x, y \in \phi_j} \|x - y\|^2 = \hat{w}_j'^T D \hat{w}_j'$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes D) \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$F = 1_N - \frac{2N}{k} I_N$$





$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\beta (\hat{w}_i^T \hat{w}_i - 1)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\beta (\hat{w}_i^T \hat{w}_i - 1)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$\hat{w}_i^T \beta G \hat{w}_i$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\beta (\hat{w}_i^T \hat{w}_i - 1)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$\hat{w}_i^T \beta G \hat{w}_i$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$G = 1_k - 2I_k$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\beta (\hat{w}_i^T \hat{w}_i - 1)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$\hat{w}_i^T \beta G \hat{w}_i$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$G = 1_k - 2I_k$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$

$$\hat{w}^T Q^T (I_N \otimes \beta G) Q \hat{w}$$



$$\alpha (\hat{w}_j'^T \hat{w}_j' - N/k)^2$$

$$\beta (\hat{w}_i^T \hat{w}_i - 1)^2$$

$$\hat{w}_j'^T \alpha F \hat{w}_j'$$

$$\hat{w}_i^T \beta G \hat{w}_i$$

$$F = 1_N - \frac{2N}{k} I_N$$

$$G = 1_k - 2I_k$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F)) \hat{w}$$

$$\hat{w}^T Q^T (I_N \otimes \beta G) Q \hat{w}$$

$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F) + Q^T (I_N \otimes \beta G) Q) \hat{w}$$



$$\min_{\hat{w}} \hat{w}^T \left( I_k \otimes (D + \alpha F) + Q^T (I_N \otimes \beta G) Q \right) \hat{w}$$

$$\alpha = \frac{\max(D)}{2(N/k) - 1}$$

$$\beta = \max(D)$$



$$\min_{\hat{w}} \hat{w}^T (I_k \otimes (D + \alpha F) + Q^T (I_N \otimes \beta G) Q) \hat{w}$$

$$\begin{aligned} \min_W & \sum_{l=1}^k \sum_{j=1}^N \sum_{i=1}^N \sum_{m=1}^d w_{il} (x_{im} - x_{jm})^2 w_{jl} \\ & + \alpha \sum_{l=1}^k \sum_{j=1}^N \sum_{i=1}^N w_{il} f_{ij} w_{jl} + \beta \sum_{l=1}^N \sum_{j=1}^k \sum_{i=1}^k w_{li} g_{ij} w_{lj} \end{aligned}$$

- Complexity  $O(N^2kd)$

**Malinen et al.**

- Complexity  $O(N^3)$

**Scikit-learn implementation**

- Complexity  $O(Nkd)$





## Algorithms used for comparisons

- **balanced quantum k-means** (case study)
- **balanced classical k-means** (authors implementation)
- **classical k-means** (scikit-learn implementation)

classical k-means is a **valid comparison**



## Adjusted Rand Index (ARI)

- compare the similarity of two partitions of a dataset
- range from  $-1$  to  $1$
- used to compare **target partitions** vs **clustering partitions**

## Total Computing time in quantum approach

$$t = t_{QUBO_{conversion}} + t_e + t_a + t_{postprocessing} \quad (1)$$



synthetic classification datasets created with *make\_classification* (Scikit-learn)

## Datasets structure

- **N** points
- **k** classes
- **1** cluster per class
- **d** features
- clusters **centered** on a  $d$ -dimensional hypercube (with side length 2.0)
- points generated from a **normal dist.** about their cluster center (std 1.0)
- each class made of  $\frac{N}{k}$  **points**



## Classical Machine

- 2.7 GHz Dual-Core Intel i5
- 8 GB 1.867 MHz DDR3 memory

## Quantum Machine

- D-Wave 2000Q quantum computer
- 2048 qubits, 5600 inter-qubit connections

## Technical Aspects

- **quantum pre/post-processing** done via the above **classical** machine
- quantum **annealing** operation **performed 100 times** for each experiment
- only ground state is used



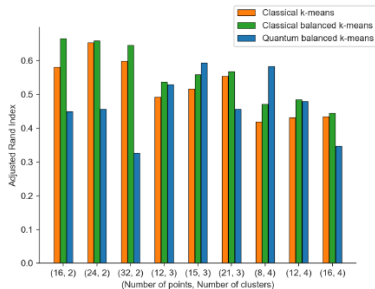
## Experiments Setup

- clustering quality of the 3 algorithms is compared
- each algorithm evaluated on different **problem types**
  - total of 9 problem types
  - defined by (*num. of points*, *num. of clusters*)
- for each problem type:
  - all the 3 algorithm evaluated on 50 **synthetic datasets**



## Commenting Results for Quantum Approach

- performances drop for  $k = 2$ 
  - less way to cluster  $\implies$  local solution is more likely to be the correct one
- performances drop as the problem size increase
  - reflection of the quantum hardware



## Limitations faced

- **Variable limitation** D-Wave 2000Q qubit limitation for problems  $Nk > 64$  var.
- **Qubit connectivity** "limitation"  $\Rightarrow$  higher embedding time

## Approximations

- Quantum run time for larger problems ( $Nk > 64$ )
  - used to evaluate scalability of the Quantum Approach
  - measure  $t_{QUBO_{conversion}}$  (measurable)
  - estimate embedding time  $t_e$  (from smaller problems)
  - estimate annealing time  $t_a$  (constant, averaging smaller problems)
  - measure  $t_{postprocessing}$



embedding algorithms chosen **scales quadratically** in the number of **binary variables** of the QUBO

$$t_e = 1.887 \times 10^{-6}(Nk)^2 + 4.632 \times 10^{-6}(Nk) + 4.022 \times 10^{-4} \quad (2)$$

$$t_a = 0.03481 \pm 0.00008 \quad (3)$$





## Experiments to assess scalability

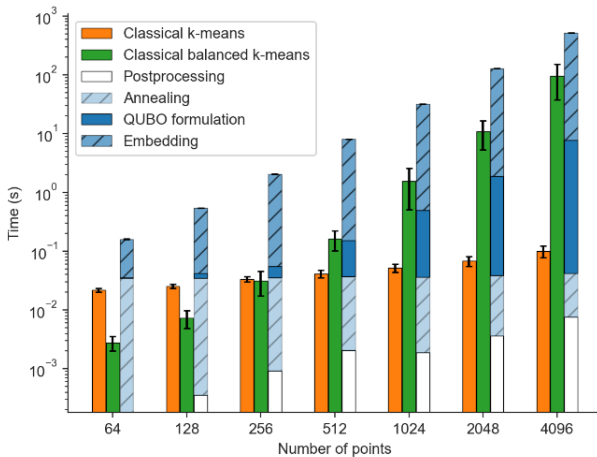
- baselines evaluated on the three variables:
  - $N$  data points
  - $k$  clusters
  - $d$  features
- $\forall$  **problem type** baselines runned on 50 **synthetic datasets**



## Setup and Considerations

- baselines evaluated on increasing **data points**
- fixed cluster  $k = 4$  and features  $d = 2$
- considerations:
  - quantum is outperformed (due to embedding time)
  - future embedding time improvements may surpass classical balanced ( $N \geq 1024$ )
  - classical k-means scales the best expected since its **time complexity**  $O(Nkd)$  vs quantum balanced  $O(N^2kd)$

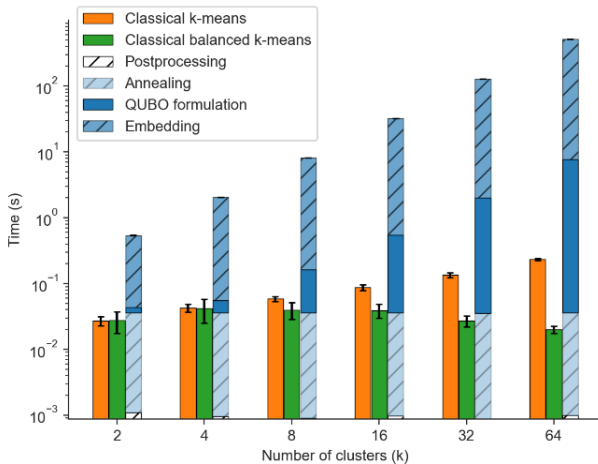




## Setup and Considerations

- baselines evaluated on increasing **cluster size**
- fixed data points  $N = 256$  and features  $d = 8$
- considerations:
  - quantum scales worse on cluster size w.r.t. to other approaches
  - expected: third term on QUBO has  $O(Nk^2)$  time complexity

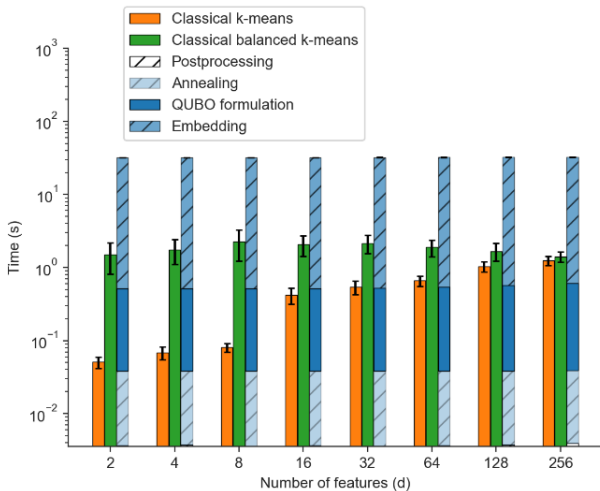




## Setup and Considerations

- baselines evaluated on increasing **features number**
- fixed data points  $N = 1024$  and cluster  $k = 4$
- considerations:
  - quantum is the worse on time
  - quantum is promising in a future perspective, depending on embedding process optimizations
  - quantum approach scales better w.r.t. to classical *k-means* on  $d$
  - expected: QUBO formulation only requires one comput. related to the dimension of the dataset
  - *classical balanced k-means* scales better in  $d$  w.r.t. to quantum approach
  - expected: *quantum balanced*  $O(N^2kd)$  vs *classical balanced*  $O(N^3)$





## The Iris Dataset

- Reduced due to qubit limitations on modern hardware
- Pick  $N/k$  points from  $2 \leq k \leq 3$  of the data set's classes

## Experiments Run

- All the 3 clustering algorithms were tested
- Experiments are run on 50 subsets of the dataset

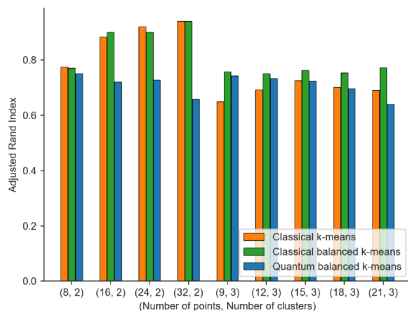
## Results

- $k = 2$ 
  - Trivial case, points are linearly separable
  - Classical algorithms perform better than quantum
  - Evident as the number of binary variables ( $Nk$ ) increases





- $k = 3$ 
  - Similar performance to **classical balanced** k-means
  - Outperforms **Scikit-Learn** implementation
  - Performance of the QA degrades as the problem size increases



- Enhancements provided by adiabatic computers for solving **NP**-Hard or **NP**-Complete problems
- Promising result for Quantum Machine Learning
- The approach targets the global solution of the training problem **better** than the classic alternatives
- The **D-Wave 2000Q** machine
- Quantum approach partitions data with similar accuracy to the classical approaches
- The approach assumes viability as the quantum hardware improves



- Bring the QUBO formulation to the generic k-means training problem

- **Balanced Problem**

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x,y \in \phi_j} \|x - y\|^2 \implies \min_{\Phi} \sum_{j=1}^k \sum_{x,y \in \phi_j} \|x - y\|^2$$

- **Generic Problem**

$$\min_{\Phi} \sum_{j=1}^k \frac{1}{2|\phi_j|} \sum_{x,y \in \phi_j} \|x - y\|^2$$

- Use elements of the approach to formulate quantum algorithms for similar clustering models
  - k-medoids clustering
  - fuzzy C-means clustering
- Cluster larger datasets



How complex is to construct the QUBO ?

$$\min_{\Phi} \sum_{j=1}^k \sum_{x,y \in \phi_j} \|x - y\|^2$$



$$\min_{\hat{w}} \hat{w}^T \left( I_k \otimes (D + \alpha F) + Q^T (I_N \otimes \beta G) Q \right) \hat{w}$$

**Complexity:**  $O(N^2kd)$

Since  $kd < N$ :

- **Better** than classical balanced k-means:  $O(N^3)$
- **Worse** than Scikit Learn implementation:  $O(Nkd)$



Hyperparameter  $\alpha$  allows to make considerations in the data preparation phase of the clustering algorithm:

- **Completely unbalanced**  $\implies$  use Scikit-Learn implementation
- **Fairly Balanced**  $\implies$  tuning on  $\alpha$  and use Quantum Balanced implementation
- **Balanced**  $\implies$  use Quantum Balanced implementation with  $\alpha < \beta$

### Tuning $\alpha$

- Modifies the curvature of the quadratic function to optimize
- By making  $\alpha$  looser we change the position of the optimum allowing to cluster datasets that are not completely balanced
- Tuning  $\alpha$  allows to prepare the algorithm on how much balanced the dataset will be



## Variables and Density of the QUBO

- In the QUBO formulation we introduce  $k$  binary variables for each variable in the original problem

$O(Nk)$  **variables**

- Efficient embedding algorithms [30] allow for a density of

$O(N^2k^2)$  **qubits**

TABLE I  
NUMBER OF BINARY VARIABLES AND AVERAGE NUMBER OF QUBITS USED IN THE QUANTUM APPROACH.

$(N, k)$	(16, 2)	(24, 2)	(32, 2)	(12, 3)	(15, 3)	(21, 3)	(8, 4)	(12, 4)	(16, 4)
Variables	32	48	64	36	45	63	32	48	64
Qubits	185	429	794	244	381	743	209	456	806

[30] P. Date, R. Patton, C. Schuman, and T. Potok, “Efficiently embedding qubo problems on adiabatic quantum computers,” Quantum Information Processing, vol. 18, no. 4, p. 117, 2019.



## Can we cluster larger datasets on Advantage?

### D-Wave 2000Q

- 2048 qubits
- 6,016 couplers
- 128,472 JJs



### Advantage

- 5640 qubits
- 40,484 couplers
- 1,030,000 JJs



**Thanks for your Attention**

---