Balanced K-Means Clustering on an Adiabatic Quantum Computer

Applied Quantum Machine Learning Project

Politecnico di Milano

July 1, 2021

Authors:
Pierriccardo Olivieri
Francesco Piro
Matteo Sacco

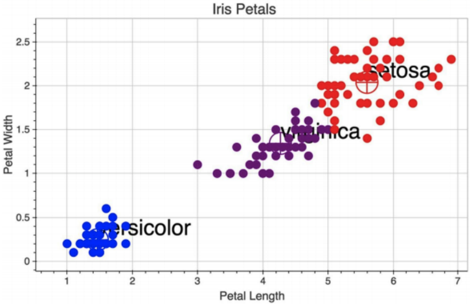
Professors:
Cremonesi Paolo
Alessandro Luongo
Maurizio Ferrari Dacrema

Outline

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

Advantages over classical approaches

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



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, "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.

QUBO formulation

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

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

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

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

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

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

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

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

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

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

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

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

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

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

Distance matrix: D

Assignment matrix: $\hat{W}$

QUBO formulation

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

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

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

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

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

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

Distance matrix: D
Assignment matrix: $\hat{W}$

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

QUBO formulation

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

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

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$$

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

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

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

Distance matrix: D

Assignment matrix: $\hat{W}$

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

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

Penalties

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

Penalties

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

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

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

Penalties

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

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

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

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

Penalties

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

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

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

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

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

Penalties

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

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

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

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

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

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

$$G = 1_k - 2I_k$$

Penalties

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

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

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

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

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

$$G = 1_k - 2I_k$$

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

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

Penalties

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

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

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

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

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

$$G = 1_k - 2I_k$$

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

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

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

Hyper-parameters tuning

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

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

Converting equation into a QUBO

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

$$\min_{W} \sum_{l=1}^{k} \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{m=1}^{d} w_{il} \left(x_{im} - x_{jm} \right)^2 w_{jl}$$

$$+\alpha \sum_{l=1}^{k} \sum_{j=1}^{N} \sum_{i=1}^{N} w_{il} f_{ij} w_{jl} + \beta \sum_{l=1}^{N} \sum_{j=1}^{k} \sum_{i=1}^{k} w_{li} g_{ij} w_{lj}$$

- Complexity $O(N^2 kd)$

Malinen et al.

- Complexity $O(N^3)$

Scikit-learn implementation

- Complexity $O(Nkd)$

Algorithms used for comparisons

- **balanced quantum k-means** $O(N^2kd)$ (case study)
- **balanced classical k-means** $O(N^3)$ (authors implementation of Malinen et al.)
- **classical k-means** $O(Nkd)$ (scikit-learn implementation Lloyd's algorithm)

classical k-means **valid comparison** (thanks to dataset structure)

Adjusted Rand Index (ARI)

- compare the similarity of two partitions of a dataset
- used to compare **ground truth labels** vs **clustering algorithm partition**
- range from -1 to 1
- ARI $= 0$ represent a random assignment

Total Computing time in quantum approach

$$t = t_{QUBO_{convertion}} + t_e + t_a + t_{postprocessing} \tag{1}$$

Synthetic datasets created with *make_classification* (Scikit-learn)

Datasets structure

- **N** points
- **k** classes
- **d** features
- each clusters **centered** on one of the **vertices** of a d-dimensional hypercube
- points generated from a **normal dist.** about their cluster center
- $\frac{N}{k}$ **exactly points** per **class**

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

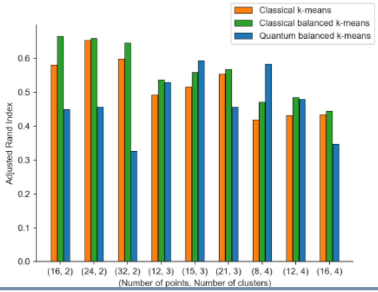
- **quantum pre/post-processing** done via **classical** machine
- quantum **anealing** operation **perfomed 100 times** for each experiment

Quality of Clustering Experiment (ARI)

- **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)*
  - e.g. *(16 points, 2 clusters)*
- for each problem type:
  - all the 3 algorithm evaluated on 50 **synthetic datasets**
  - **Averaged ARI** is reported

Commenting Results for Quantum Approach

observation	possible cause/motivation
drop in performance for $k = 2$ classical vs quantum	less way to cluster, local solution is more likely to be the correct one
quantum performances drop as the problem size increase	reflection of the quantum hardware
best quantum performances obtained for problem types (8,4) (12, 3) (12, 4)	as quantum computer improves author's approach may outperform classical



Limitations and Approximations I

Limitations faced

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

Approximations

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

Limitations and Approximations II

t_e **scale 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

- evaluation metric used is **average total computing** time (**estimated** for quantum)
- baselines evaluated on the **three variables**:
  - N data points
  - k clusters
  - d features
- baselines runned on 50 **synthetic datasets**

Scalability w.r.t. points (N) I

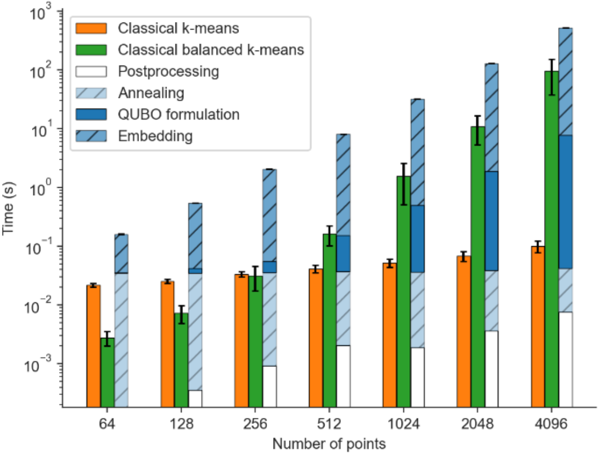
Details

- baselines evaluated on increasing **data points**
- fixed cluster $k = 4$ and features $d = 2$

Considerations

observation	motivation/improvement
quantum approach slower than both classical	dominated by embedding time, in future can improve for N>1024
classical k-means scale the best	$O(Nkd)$ vs $O(N^3)$ and $O(N^2kd)$

Scalability w.r.t. points (N) II



Scalability w.r.t. clusters (k) I

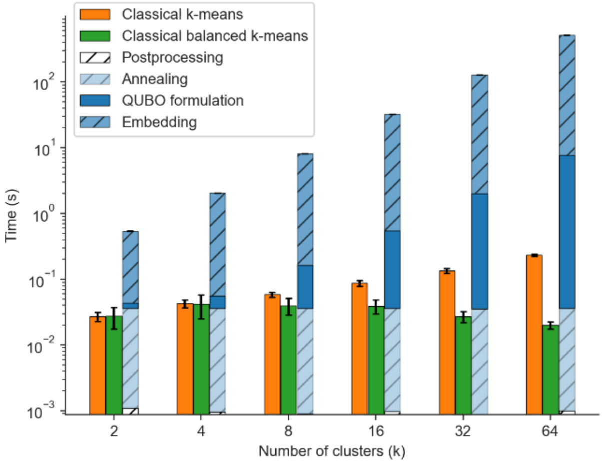
Details

- baselines evaluated on increasing **cluster size**
- fixed data points $N = 256$ and features $d = 8$

Considerations

observation	**motivation/improvement**
quantum longer **time** on k **scale worse** as k increase	Third term in the QUBO formulation time complexity $O(Nk^2)$

Scalability w.r.t. clusters (k) II



Scalability w.r.t. features (d) I

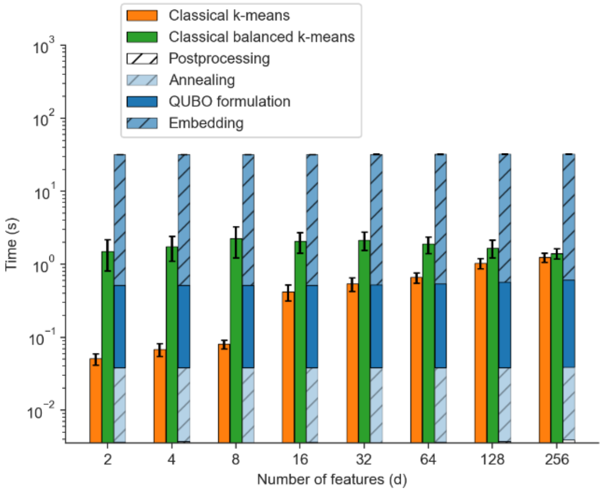
Details

- baselines evaluated on increasing **features number**
- fixed data points $N = 1024$ and cluster $k = 4$

Considerations

observation	**motivations/improvements**
quantum had longest time	**improvements** in hardware and embeddings could outperform classical k-means $d > 128$ classical balanced k-means $d < 256$
quantum scales better w.r.t. classical k-means as d increases	QUBO formulation only requires one comput. of the distance matrix classical re-compute distance from centroids at each iteration
classical balanced k-means scales better in d w.r.t. to quantum approach	$O(N^2kd)$ vs $O(N^3)$

Scalability w.r.t. features (d) II



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

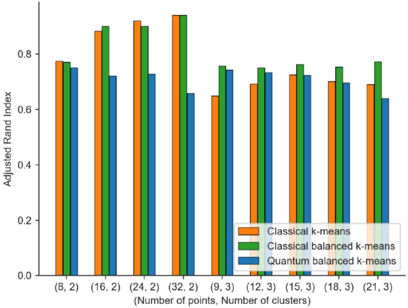
- All the 3 clustering algorithms were tested
- Experiments are run on 50 subsets of the dataset

Results

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

Clustering a Benchmark Data Set II

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



Author's Conclusions

- 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

Future Works

- 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 ?

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

$$\Downarrow$$

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

Complexity: $O(N^2 kd)$

Since $kd < N$:

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

Critical View II

Hyperparameter α 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 α and use Quantum Balanced implementation
- **Balanced** $\implies$ use Quantum Balanced implementation with $\alpha < \beta$

Tuning α

- Modifies the curvature of the quadratic function to optimize
- By making α looser we change the position of the optimum allowing to cluster datasets that are not completely balanced
- Tuning α 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

$$O(Nk)\ \textbf{variables}$$

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

$$O(N^2k^2)\ \textbf{qubits}$$

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