Data Clustering as a Quantum Computing Use-Case

Gerhard Hellstern
DHBW Stuttgart
Stuttgart, Germany
gerhard.hellstern@dhbw-stuttgart.de

Martin Zaefferer
DHBW Ravensburg
Ravensburg, Germany
zaefferer@dhbw-ravensburg.de

***Extended Abstract*— The research project QORA and QORA II, funded by the Ministry of Economics, Labor and Tourism Baden-Württemberg, explored in the last four years potential applications of Quantum Computing (QC) with resilient algorithms. In particular use cases in the domain of finance were investigated: portfolio optimization [1] and feature selection for a credit scoring algorithm [2]. To gain a business impact, potential quantum methods are compared to established methods of classical, i.e. non-quantum, computing.**

One further potential application in the finance domain is data clustering within an unsupervised learning approach. Data clustering methods are regularly applied in banking and insurance companies to put customers into different segments. Then, for each segment, a specific and targeted marketing campaign can be applied.

A popular classical formulation of the clustering problem is using the K-means approach, which is known to be NP-hard [3]. The corresponding K-means algorithm, due to its greedy nature, has no guarantees to find the global optimum of a given problem. Quantum computers might provide an improvement due to their generic characteristics – superposition and entanglement. The improvement may arise in terms of accuracy and efficiency of the solution. To this end, the clustering problem is transformed into a quadratic binary optimization (QUBO) problem, which can be solved using QC algorithms like QAOA, c.f., e.g. [1,2].

In our study, we are particularly interested in the scaling behavior of the quantum algorithm QAOA compared with its classical counterparts. As generic data, we use the "moons" and "blob" data sets, which are well-known clustering benchmarks in the machine-learning community. When using QAOA in this setting, one data-point is mapped to one quantum bit (qubit). Carefully analyzing the resources needed to solve problem sizes up to twenty qubits allows us to extrapolate, how the runtime and solution quality scale with increasing number of data samples (i.e., more bits/qubits). Here, we are working in a regime of up to one hundred qubits.

In a next step we compare the scaling of QAOA with well-established optimization methods. These classical approaches for solving the QUBO-problem can be applied out of the box to a regime of up to one hundred data points and the scaling behavior can be extracted directly. We investigate three classical solvers: CPLEX, Gurobi, and the Goemans-Williamson method. We observe that the scaling is sensitive to the properties of the data and differs significantly between the algorithms. This contrasts with QAOA, where the scaling is rather insensitive to the data. Based on that observation, we generated additional data sets that are harder to solve for the classical solvers, and potentially not as hard for QAOA.

A further topic of our research is to solve the QAOA problem with tensor networks of varying bond dimension. Again, the data dependence of the scaling behavior as well as the accuracy of the solution is analyzed.

Our study and its results can help to clarify the discussion of whether a quantum advantage could be expected for optimization problems, how big that advantage might be, and how many qubits it would necessitate. While our results are done on quantum simulators, the latest results from the quantum hardware research stipulate, that an amount of ca. one hundred error-free qubits might be available in the near future [4,5].

Keywords*— *Machine Learning, Quantum Computing, Unsupervised Learning, Clustering, Optimization

[1] Brandhofer, S.; Braun, D.; Dehn, V.; Hellstern G.; Hüls,M.; Ji, Y.; Polian, I.; Bhatia, A.S.; Wellens, T.; Benchmarking the performance of portfolio optimization with QAOA. Quantum Inf Process 22, 25 (2023). https://doi.org/10.1007/s11128-022-03766-5

[2] Hellstern, G.; Dehn, V.; Zaefferer, M.;"Quantum computer based feature selection in machine learning" IET Quantum Communication, (2024) https://doi.org/10.1049/qtc2.12086,

[3] Meena, M.; Nimbhorkar, P.; and Varadarajan, K.; The planar k-means problem is NP-hard; Theoretical Computer Science 442 (2012)

[4] Bluvstein, D., Evered, S.J., Geim, A.A. *et al.* Logical quantum processor based on reconfigurable atom arrays. *Nature* 626, 58–65 (2024). https://doi.org/10.1038/s41586-023-06927-3

[5] Bravyi, S., Cross, A.W., Gambetta, J.M. et al. High-threshold and low-overhead fault-tolerant quantum memory. Nature 627, 778–782 (2024) https://doi.org/10.1038/s41586-024-07107-7