ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Quantum Feature Selection with Variance Estimation

Alessandro Poggiali, Anna Bernasconi, Alessandro Berti, Gianna M. Del Corso, and Riccardo Guidotti *

Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3 56127 Pisa - Italy

Abstract. The promise of quantum computation to achieve a speedup over classical computation led to a surge of interest in exploring new quantum algorithms for data analysis problems. Feature Selection, a technique that selects the most relevant features from a dataset, is a critical step in data analysis. With several Quantum Feature Selection techniques proposed in the literature, this study exhibits the potential of quantum algorithms to enhance Feature Selection and other tasks that leverage the variance. This study proposes a novel quantum algorithm for estimating the variance over a set of real data. Importantly, after state preparation, the algorithm's complexity exhibits logarithmic characteristics in both its width and depth. The quantum algorithm applies to the Feature Selection problem by designing a Hybrid Quantum Feature Selection (HQFS) algorithm. This work showcases an implementation of HQFS and assesses it on two synthetic datasets and a real dataset.

1 Introduction

With quantum computation promising a speedup over classical computation, exploring new quantum algorithms for known problems becomes crucial. In this work, we consider the *Feature Selection* problem [1], a data analysis technique that identifies and selects the most relevant features from a dataset. This technique minimizes computational costs by discarding the less relevant features when creating models from a given dataset.

Several Quantum Feature Selection algorithms have been recently proposed in the literature [2–4]. Among the Feature Selection techniques [5], here we focus on an unsupervised Feature Selection method, i.e., a method that does not use explicit target labels or class information to select the relevant features. In particular, we consider a Feature Selection technique for numerical variables based on removing features whose variance is below a given threshold [6]. More precisely, we propose a novel quantum algorithm (QVAR) for estimating the variance over a given set of values. Then we apply it to the problem of Feature

*This study was carried out within the National Centre on HPC, Big Data and Quantum Computing - SPOKE 10 (Quantum Computing) and received funding from the European Union Next-GenerationEU - National Recovery and Resilience Plan (NRRP) - MISSION 4, COMPONENT 2 - CUP N. I53C22000690001. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. The work has been partially supported also by INdAM-GNCS Project CUP_E53C22001930001

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

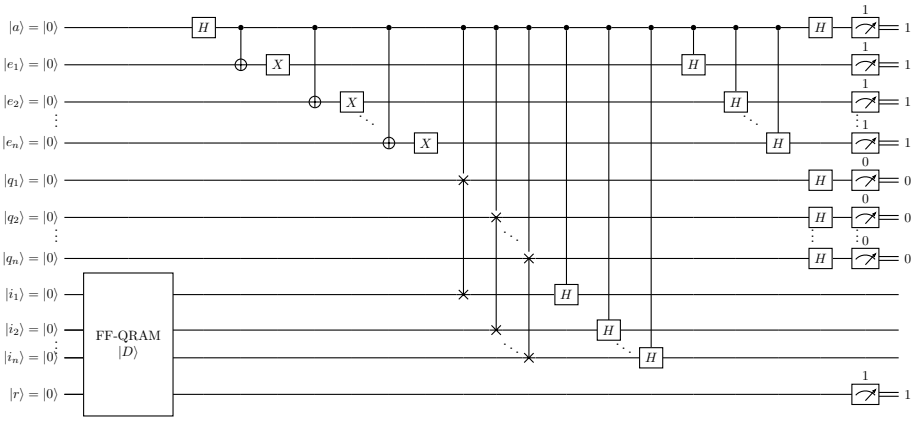

Fig. 1: Quantum circuit for variance estimation.

Selection by designing a Hybrid Quantum Feature Selection (HQFS) algorithm based on the variance filter [7]. Eventually, we showcase an implementation of the HQFS algorithm, assessing it on two synthetic datasets and a real dataset.

The work is organized as follows. Section 2 describes the quantum variance estimation algorithm, while in Section 3, we show our quantum Feature Selection algorithm. Section 4 reports experimental simulations. Finally, Section 5 summarizes the results obtained and proposes possible future works.

2 Quantum Algorithm for Variance Estimation

Given a set $D = \{d_0, d_1, \ldots, d_{N-1}\}$ of N real values, we encode D into a quantum state $|D\rangle$ by means of FF-QRAM [8], one of the currently available state preparation techniques. In particular, FF-QRAM encodes the values $d_t \in D$ in the amplitude of a one-qubit register $|r\rangle$, and employs $n = \log_2 N$ additional qubits to index each value. We design the algorithm for the variance estimation according to the following formula for the variance: $\sigma^2 = \frac{\sum_{t=0}^{N-1}(d_t - \mu_D)^2}{N}$, where $\mu_D = \frac{1}{N}\sum_{k=0}^{N-1} d_k$ is the mean of the values in D.

This algorithm requires in total $3n + 2$ qubits. Figure 1 illustrates the quantum circuit implementing the variance estimation algorithm. Due to space limitations, the formal proof of correctness will be given and discussed in the extended version of this paper. Here, we only provide the key steps involved in the algorithm. Let us prepare an ancilla qubit $|a\rangle$, a register qubit $|r\rangle$, and three quantum registers $|i\rangle^{\otimes n}$, $|e\rangle^{\otimes n}$, and $|q\rangle^{\otimes n}$ in the following quantum state: $|\psi_0\rangle = |0\rangle_a |0\rangle_e^{\otimes n} |0\rangle_q^{\otimes n} |0\rangle_i^{\otimes n} |0\rangle_r$. Then, we encode D into a quantum state $|D\rangle$ by means of FF-QRAM:

$$|\psi_1\rangle = \frac{1}{\sqrt{2^n}} |0\rangle_a |0\rangle_e^{\otimes n} |0\rangle_q^{\otimes n} \sum_{t=0}^{N-1} |t\rangle_i |d_t\rangle_r ,$$

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

where $|d_t\rangle$ corresponds to the quantum state encoding the value d_t [9]. We create an equal superposition with the first H gate on $|a\rangle$. Then, we compute the mean μ_D in the branch where $|a\rangle = 1$. In the end, we apply another H gate on $|a\rangle$ to cause each $|d_t\rangle_r$ stored in the branch where $|a\rangle = 0$ to collide with the mean μ_D. Finally, the H gates on $|q\rangle^{\otimes n}$ result in the actual sums between each d_t. Before measurement, the final state considering the configurations where $|ae^{\otimes n}q^{\otimes n}\rangle = |11^{\otimes n}0^{\otimes n}\rangle$ is:

$$\frac{1}{2\sqrt{N}}|1\rangle_a|1\rangle_e^{\otimes n}|0\rangle_q^{\otimes n}\left(\frac{1}{\sqrt{N}}\sum_{t=0}^{N-1}|t\rangle_i|d_t\rangle_r - \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1}|t\rangle_i\frac{1}{N}\sum_{k=0}^{N-1}|d_k\rangle_r\right)$$
$$= \frac{1}{2N}|1\rangle_a|1\rangle_e^{\otimes n}|0\rangle_q^{\otimes n}\sum_{t=0}^{N-1}|t\rangle_i\left(|d_t\rangle_r - \frac{1}{N}\sum_{k=0}^{N-1}|d_k\rangle_r\right).$$

Eventually, we perform a post-selective measurement on qubits $|ae^{\otimes n}q^{\otimes n}r\rangle = |11^{\otimes n}0^{\otimes n}1\rangle$, which succeeds with probability

$$P_{succ} = \frac{1}{4N^2}\sum_{t=0}^{N-1}(d_t - \frac{1}{N}\sum_{k=0}^{N-1}d_k)^2, \tag{1}$$

which corresponds to the variance of D, divided by $4N$. The cost of the actual computation of the variance using the circuit in Figure 1 is $O(\log N)$ both in terms of width and depth, excluding the cost of data encoding. However, since the success probability corresponds to the variance of D when the variance is low, the success probability of the required post-selection step is also low. Moreover, since the variance derives from a probability and therefore depends on the amplitudes of the final quantum system, the number of circuit repetitions (shots) required to accurately estimate the variance should be large. In particular, from the Chebyshev inequality follows that the number R of shots needed to derive a quantum estimate of the variance within an error ϵ from the classical one grows with $O(\epsilon^{-2})$ with probability $(1-\delta)$, where $\delta = \frac{P_{succ}(1-P_{succ})}{R\epsilon^2}$.

An effective solution to this problem is to use the Amplitude Estimation (AE) [10] subroutine, which easily addresses the issue. We call QVAR the algorithm which exploits AE to encode the amplitude of the target configuration over m additional qubits in the computational basis, where $m = O(\log\frac{1}{\epsilon})$. This value of m sets the precision of the estimation. In our case, we employ AE to estimate the amplitude of the target configuration $|ae^{\otimes n}q^{\otimes n}r\rangle = |11^{\otimes n}0^{\otimes n}1\rangle$ by measuring the m additional qubits. The measurement output represents an approximation of the variance in the computational basis. In general, the depth of AE is $O(\delta\frac{1}{\epsilon} + \log\log\frac{1}{\epsilon})$ [11] where δ is the depth of the oracle. In our case, $\delta = O(\log N)$, assuming the existence of an efficient quantum state preparation technique. Thus, the overall complexity of QVAR is $O(\log N)$, and the number of qubits is $m + \log_2 N$. We can increase the precision of QVAR by interpolating the measurement probabilities and computing the maximum likelihood estimator (we call this method ML-QVAR) [12].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

3 Hybrid Quantum Feature Selection

This section presents the Hybrid Quantum Feature Selection (HQFS) algorithm, which exploits the variance estimation techniques, QVAR and ML-QVAR.

Setting the stage. Given a multidimensional dataset $D \in \mathbb{R}^{N \times M}$ with N records and M features, we denote as $D_{i,j}$ the j-th feature value of the i-th record. The HQFS algorithm iterates over the features and computes the variances $\sigma_j^2 = variance(d_{0,j}, \cdots, d_{N-1,j})$ for $j \in [0, \cdots, M-1]$. HQFS employs a distinct QVAR circuit for each feature j to calculate each variance. In total, HQFS uses M QVAR circuits, one for each feature. Then, HQFS drops all features whose variance is below a given threshold t.

Algorithm 1 HQFS

Input: D - input dataset, t - variance threshold
Output: F - list of selected features

$F \leftarrow [f_0, \cdots, f_{M-1}]$ // list of selected features initially containing all features
for $j \in [0, \cdots M-1]$ **do** // classically iterate among all feature
  $\sigma_j^2 \leftarrow QVAR(D_{:,j})$; // compute the variance using the QVAR algorithm.
  if $\sigma_j^2 \leq t$ **then** // check if feature f_j is uninformative
    $F \leftarrow F \setminus \{f_j\}$; // remove feature f_j to the list of selected features
return F ; // return the list of selected features

We can use either the QVAR or ML-QVAR as a variance estimation method. The complexity of Algorithm 1 is $O(M \log N)$ if we assume available efficient methods for reconstructing the initial state. Using the Qiskit framework by IBM, we implement two versions of HQFS based on the variance estimation methods provided, HQFS and ML-HQFS, respectively[1].

4 Experiments

We conduct the experiments using the QASM SIMULATOR of Qiskit, which simulates the quantum circuits using classical hardware. Before assessing the HQFS algorithm, we compare the QVAR and ML-QVAR algorithms with the classical variance on 5 sets of 8 random uniform values in the range [-1,1]. In Figure 2, we plot the Mean Squared Error (MSE) with respect to the classical variance, varying the number m of additional qubits.

We then evaluate the performance of HQFS and ML-HQFS on two synthetic datasets and a real dataset by varying the parameter m. We assess our results by comparing the similarities of the final ranking of features of HQFS and ML-HQFS with respect to the ranking from the classical variance. To measure the similarity between the rankings, we use the Rank Biased Overlap (RBO) measure [13], which weights the ranking similarity on the top ranks and takes value in [0,1]. A high value of RBO corresponds to a high similarity on the top ranks.

[1] Code publicly available at `https://github.com/AlessandroPoggiali/HQFS`

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

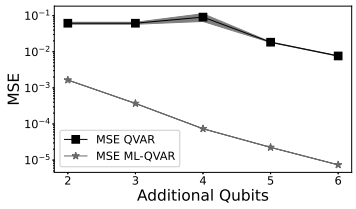

Fig. 2: MSE for QVAR and ML-QVAR with respect to the classical variance.

Dataset. For the experimental evaluation of HQFS, we consider two synthetic datasets (`synth_1` and `synth_2`), and the real dataset `wine`. Both synthetic datasets have $N = 32$ records and $M = 10$ features: 7 informative features with high variance and 3 uninformative features with low variance, which are not expected to provide useful information for the analysis. The informative features are sampled from uniform distributions in [-1,1], while the uninformative features are sampled from two normal distributions with low variance. In particular, the uninformative features for `synth_1` are sampled from a normal distribution with a standard deviation of 0.05. In contrast, the uninformative features for `synth_2` are sampled from a normal distribution with a standard deviation of 0.5.

synth_1		
m	HQFS	ML-HQFS
2	0.05	0.37
3	0.05	0.35
4	0.05	0.37
5	0.15	0.99
6	0.43	0.98
synth_2		
m	HQFS	ML-HQFS
2	0.21	0.20
3	0.21	0.20
4	0.21	0.20
5	0.27	0.20
6	0.31	0.37

Table 1: RBO measures on synthetic datasets.

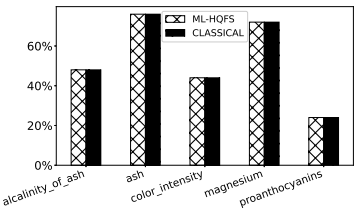

Fig. 3: Features selected by ML-HQFS on `wine` dataset.

Results. In Table 1, we report the HQFS and ML-HQFS results on `synth_1` and `synth_2` datasets. To mitigate the high computational cost of simulation, we test the algorithm with a maximum value of $m = 6$. However, theoretical analysis suggests that increasing the value of m can yield superior results. We note that ML-HQFS performs better than HQFS for `synth_1` datasets. However, the RBO measure is unsuitable for the `synth_2` dataset because the variance of the uninformative features is high, and then the algorithm fails to distinguish between informative and uninformative features. If we are only interested in recognizing uninformative features, we see that ML-HQFS always gets an accuracy of 100% while the accuracy of HQFS ranges from 30% to 80% depending on the number m of additional qubits. In Figure 3 we show how the ML-HQFS algorithm (with $m = 6$) behaves on the real dataset `wine`, sampling randomly

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

16 records 25 times. The figure shows that ML-HQFS selects the same feature as the classical algorithm.

5 Conclusion

In this paper, we proposed a novel quantum algorithm (QVAR) for estimating the variance of a superposition of values. As a use case, we designed an efficient Hybrid Quantum Feature Selection (HQFS) algorithm that exploits the quantum variance estimation through the Amplitude Estimation subroutine. The experiments have shown that QVAR outputs a good estimate of the classical variance if the number of additional qubits is properly chosen. Therefore, the final ranking of features produced by HQFS is similar to the ranking produced by the classical algorithm, especially considering low-variance features, meaning that HQFS correctly eliminates uninformative features. Possible feature works include applying the QVAR algorithm to other tasks that leverage the variance.

References

[1] J. Li, K. Cheng, S. Wang, F. Morstatter, RP. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[2] S. Mücke, R. Heese, S. Müller, M. Wolter, and N. Piatkowski. Quantum feature selection. *arXiv:2203.13261*, 2022.

[3] R. Nembrini, M. Ferrari Dacrema, and P. Cremonesi. Feature selection for recommender systems with quantum computing. *Entropy*, 23(8):970, 2021.

[4] S. Chakraborty, SH. Shaikh, A. Chakrabarti, and R. Ghosh. A hybrid quantum feature selection algorithm using a quantum inspired graph theoretic approach. *Applied Intelligence*, 50(6):1775–1793, 2020.

[5] V. Kumar and S. Minz. Feature selection: a literature review. *SmartCR*, 4(3):211–229, 2014.

[6] PN. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[7] S. Solorio-Fernández, JA. Carrasco-Ochoa, and JF. Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948, 2020.

[8] DK. Park, F. Petruccione, and JKK. Rhee. Circuit-based quantum random access memory for classical data. *Scientific reports*, 9(1):3949, 2019.

[9] M. Schuld and F. Petruccione. *Supervised learning with quantum computers*, volume 17. Springer, 2018.

[10] G. Brassard, P. Hoyer, M. Mosca, and A. Tapp. Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305:53–74, 2002.

[11] T. Giurgica-Tiron, I. Kerenidis, F. Labib, A. Prakash, and W. Zeng. Low depth algorithms for quantum amplitude estimation. *Quantum*, 6:745, 2022.

[12] D. Grinko, J. Gacon, C. Zoufal, and S. Woerner. Iterative quantum amplitude estimation. *npj Quantum Information*, 7(1):52, 2021.

[13] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.