ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Quantum Annealing based Feature Selection

Daniel Pranjić, Bharadwaj Chowdary Mummaneni and Christian Tutschku *

Fraunhofer Institute for Industrial Engineering - Nobelstr. 12, 70569 Stuttgart - Germany

Abstract. Feature selection is crucial for enhancing the accuracy and efficiency of machine learning models. Calculating the optimal feature set for maximum mutual information (MI) and conditional mutual information (CMI) remains computationally intractable for large datasets on classical computers, even with approximation methods. This study employs a Mutual Information Quadratic Unconstrained Binary Optimization (MIQUBO) formulation, enabling its solution on a quantum annealer. To showcase its real-world applicability, we apply MIQUBO to forecasting the price of used excavators. Our results indicate that using the MIQUBO approach there might be an improvement in the prediction of machine learning models for datasets, with a smaller MI concentration.

1 Introduction

Quantum machine learning (QML) [1] is a rapidly evolving field investigating the intersection of quantum computing and machine learning algorithms. It aims to leverage the unique properties of quantum mechanics, such as superposition and entanglement, to enhance the capabilities of traditional machine learning (ML) methods. Feature selection helps create simpler models, which in turn reduces computational demands. The high dimensionality of most datasets mandates the development of efficient and effective feature selection algorithms. Classical ML methods have been extensively applied to calculate the residual value of construction equipment [2]. Both conventional ML and automatized ML (AutoML) methods were shown to yield good results for different applications and datasets [3]. In the context of forecasting the residual values of used excavators Support Vector Machines (SVMs) and hybrid quantum SVMs have been applied [4], where quantum SVMs were shown to be competitive with classical SVMs.

In this work we demonstrate that the best feature combinations can be obtained for real world problems by using a hybrid approach that uses classical methods together with quantum annealing for solving QUBOs. While writing this paper, the feature selection in machine learning based on the mutual information QUBO (MIQUBO) was addressed in [5] where the problem was tackled with universal quantum computers using the variational quantum eigensolver and quantum approximate optimization algorithms.

*This work was funded by the German Federal Ministry of Economic Affairs and Climate Action in the research project AutoQML (grant no. 01MQ22002A). The authors thank Horst Stühler for providing the Caterpillar datasets.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

2 Feature Selection based on Mutual Information

The dataset *Caterpillar (only model 308)*, has 227 samples and includes categorical features. We convert the categories into a binary representation by using one-hot-encoding. This increases the dimensionality from 5 to 27 features. As a reference, we also analyze the larger and more higher-dimensional *Caterpillar (all models)* dataset, which consists of 2996 samples with 6 features out of which 3 are categorical. After one-hot-encoding the dimensionality increases from 6 to 67. These datasets were already analyzed in [4], where conventional autoencoders were used to obtain a latent space representation.

The mutual information between two random variables $x \in X, y \in Y$ quantifies how much information of x can be obtained from observations of y (and vice versa) and is defined by

$$\mathrm{MI}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{1}$$

where $p(x,y)$ is the joint probability of the marginal probabilites $p(x)$ and $p(y)$. Using the Shannon entropy $S(X) = -\sum_{x\in X} p(x)\log p(x)$, we can rewrite Eq. (1) as $\mathrm{MI}(X;Y) = S(Y) - S(Y|X)$. Additionally, we are interested in how much the target X (here: the price) is dependent on a given feature Y given the selection of another feature Z. For this, we can use conditional mutual information (CMI)

$$\mathrm{MI}(X;Y|Z) = S(X|Z) - S(X|Y;Z)\,. \tag{2}$$

where, $S(X|Y;Z)$ is

$$S(X|Y;Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\left(\frac{p(x,z|y)}{p(x|y)p(z|y)}\right), \tag{3}$$

$p(x,z|y)$ is the joint conditional probability of x and z given y. Eq. (2) can be used to determine the most optimal feature combination by selecting the k independent features $X_1,\dots X_k$ of total n features and a target variable Y that maximize the sum of all $\mathrm{MI}(X_i;Y)$ and $\mathrm{MI}(X_j;Y|X_i)$. However, this approach is limited by the growth of $\binom{n}{k}$.

Instead of calculating Eq. (2) exactly, several approximations were proposed [6, 7]. Assuming that there is a conditional independence of features F_i, F_j meaning that $p(F_i|F_j;F_k) = p(F_i|F_k)\,, \forall i,j$ and restricting Eq. (2) to three features, the optimal combination of features F is approximated by the solution of the following optimization problem

$$\underset{F}{\mathrm{argmax}} \sum_{i\in F}\left\{\mathrm{MI}(X_i;Y) + \sum_{j\in F, j\neq i}\mathrm{MI}(X_j;Y|X_i)\right\}. \tag{4}$$

This is still a NP-hard problem, but it is possible to reformulate it as a QUBO that can be solved on a quantum annealer [8].

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

3 Mutual Information QUBO on Quantum Annealers

Quantum annealing (QA) [9] is an optimization process that utilizes quantum fluctuations to find the global minimum of a given objective function over a set of candidate solutions. The time evolution of the state of a quantum system $|\psi(t)\rangle$ is described by the Schrödinger Equation

$$\mathrm{i}\frac{d\,|\psi(t)\rangle}{dt} = H(t)\,|\psi(t)\rangle\ , \tag{5}$$

where i is the imaginary unit and $H(t)$ the system's Hamiltonian. The system will remain close to the instantaneous ground state of the time-dependent Hamiltonian if the system changes sufficiently slowly by following the adiabatic condition [9]. During the adiabatic evolution, the Hamiltonian gradually transitions from the initial Hamiltonian H_I to the final Hamiltonian $H = s(t)H_I + (1-s(t))H_F$, where $s(t)$ is a function modeling the transition, such that $s(t_I = 0) = 1$ and $s(t_F) = 0$ after a certain elapsed time t_F. In this context, the ground state of the final Hamiltonian H_F encodes the solution to the problem. Notable examples are binary quadratic models (BQM), which include the Ising model and its computer science counterpart, the quadratic unconstrained binary optimization (QUBO) problem.

In the following, we show the steps to reformulate Eq. (4) into a QUBO. Given N binary variables $x_1, ..., x_N$, the quadratic formulation is written as

$$\min_{x_i, x_j} \left\{ \sum_{i}^{N} q_i x_i + \sum_{i<j}^{N} q_{i,j} x_i x_j \right\}\ , \tag{6}$$

where q_i and $q_{i,j}$ are the linear and quadratic coupling coefficients respectively. The restriction that led to Eq. (4) for MI-based feature selection naturally lends itself to being reformulated as a QUBO. Each selection of $\binom{n}{k}$ features can be represented as the bitstring $x_1, ..., x_N$ by encoding $x_i = 1$ if feature X_i should be selected, and $x_i = 0$ if not. With solutions encoded in this manner, the QUBO can be represented as $\mathbf{x}^T\mathbf{Q}\mathbf{x}$, where $\mathbf{Q}$ is a n x n matrix and $\mathbf{x}$ is a n x 1 matrix (a binary vector) that has k ones representing the k selected features. To map Eq. (4) to a QUBO, we set the elements of $\mathbf{Q}$ such that $Q_{ii} \mapsto -\mathrm{MI}(X_i; Y)$ and $Q_{i,j} \mapsto -\mathrm{MI}(X_j; Y|X_i)$. These QUBO terms are negative because the quantum computer minimizes the QUBO problem in Eq. (6), while the feature selection optimization problem in Eq. (4) maximizes the CMI.
The hybrid sampler runs three algorithms in parallel: Tabu search, simulated annealing, and QPU sub-problem sampling. Simulated Annealing is used to escape the local minima and find an approximate global minimum. In this work, 95 physical qubits were used for QPU sampling. For the annealing schedule the following parameters were chosen: sampling time 286.128 μs, anneal time per sample 600 μs and readout time per sample 95 μs. The influence of each schedule parameter was analyzed in detail in Ref. [10].

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

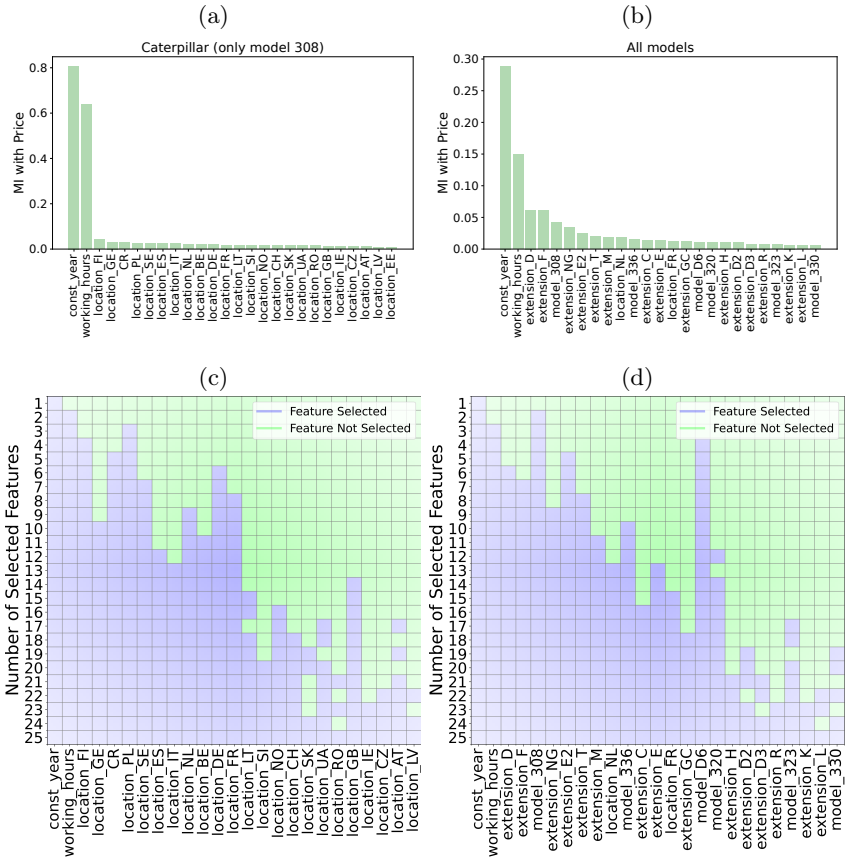

Fig. 1: Features of the a) *Caterpillar (only model 308)* & b) *Caterpillar (all models)* dataset are shown with the most mutual information (MI) between an individual input feature and the target output (here: price). c) & d) show the combination of features that maximizes Eq. (4) for a given number k of selected features for the c) *Caterpillar (only model 308)* & d) *Caterpillar (all models)* datasets.

4 Numerical Results

Fig. 1a shows the features with the highest MI of the *Caterpillar (only model 308)* dataset. It reflects that the construction year and the working hours of a used excavator are the features that carry the most mutual information about its price. We refer to this as an example of a *MI-concentrated* dataset. The information contained in the *location* feature is distributed into many individual features $\{location\_FI, location\_PL, \dots\}$ due to one-hot-encoding. In contrast,

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933.
Available from http://www.i6doc.com/en/.

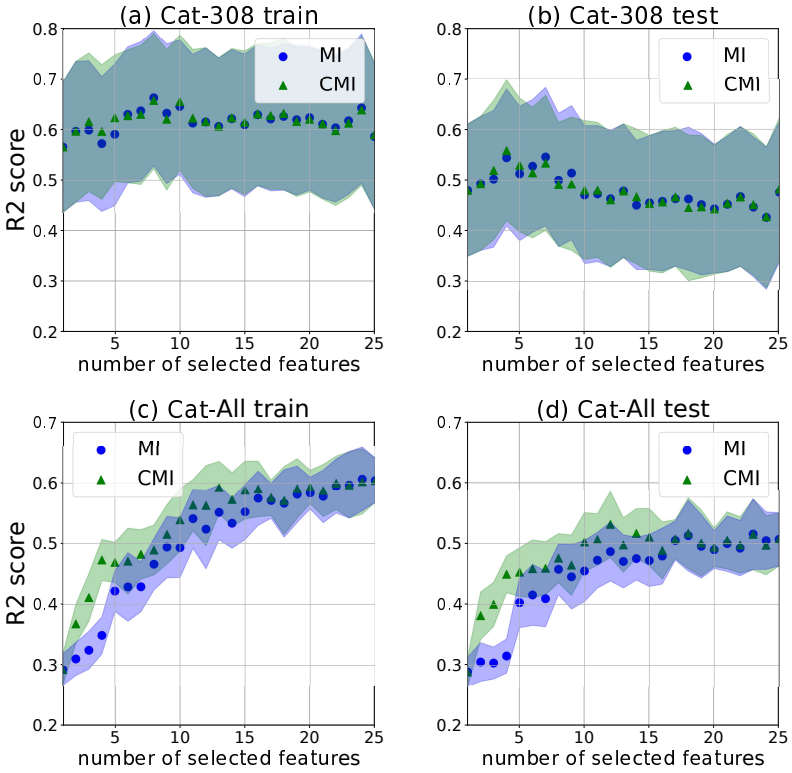

Fig. 2: Mean R2 scores of the rbf-SVMs evaluated on 100/15 randomly generated splits of *Caterpillar (only model 308)*/*Caterpillar (all models)*. Parameters: $\gamma = C = 1, \varepsilon = 10^{-3}$.

for the *Caterpillar (all models)* dataset, the MI is more equally distributed over the features and is hence a less MI-concentrated dataset.
Fig. 1c and Fig. 1d show for a given number k the best combination of features that maximizes the combined MI and CMI among the selected features. The selected features differ for both approaches.

In the following, we demonstrate empirically that a set of k selected features with maximized the sum of MI and CMI performs better in a typical machine learning model, than one that maximizes only the total MI instead. In Fig. 2 we evaluated the mean R2 score of a rbf-SVM model for 100 train-test-splits on the *Caterpillar (only model 308)* (see Fig. 2a and 2b) and 15 train-test-splits

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

on the *Caterpillar (all models)* dataset (see Fig. 2c and 2d). The data in both cases was standardized and centralized before the kernel evaluation, while we fixed $\gamma = C = 1$ and $\varepsilon = 10^{-3}$. For Fig. 2a and 2b there is no visible difference between the method of maximizing MI in comparison to maximizing the CMI. This is due to the fact that *Caterpillar (only model 308)* is a dataset with more MI-concentration than *Caterpillar (all models)* and hence in Fig. 2c and 2d, where the MI-concentration is much less, a gap between both feature selection methods opens. The difference becomes most evident for $k \leq 5$. Eventually, the difference between the information content for the set selected by MI and CMI vanishes as both methods include most of the highly informative features, reducing the performance gap between them.

5 Conclusion

This study demonstrated the potential of hybrid quantum annealing approaches for feature selection based on conditional mutual information in machine learning with MIQUBOs. We applied the obtained combinations to a real-world scenario: predicting excavator prices. These results demonstrate that for datasets with less MI-concentration the set of features that maximizes the sum of MI and CMI leads to a better performance of machine learning models. Further research can investigate the application of MIQUBO to different datasets and machine learning tasks. It would be interesting to identify the performance gap in other datasets that have even less MI-concentration.

References

[1] M. Schuld and N. Killoran, Quantum Machine Learning in Feature Hilbert Spaces, *Phys. Rev. Lett.*, **122**, 040504, 2019.

[2] H. Stühler et al., Benchmarking Automated Machine Learning Methods for Price Forecasting Applications, *arXiv preprint*, arXiv:2304.14735v1, 2023.

[3] M.A. Zöller and M.F. Huber, Benchmark and survey of automated machine learning frameworks, *Journal of artificial intelligence research*, **70**, 409-472, 2021.

[4] H. Stühler, D. Pranjić and C. Tutschku, Evaluating Quantum Support Vector Regression Methods for Price Forecasting Applications, *ICAART* **(3)**, 376-384, 2024.

[5] G. Hellstern, V. Dehn and M. Zaefferer, Quantum computer based feature selection in machine learning, *IET Quantum Communications*, **5(3)**, 232-252, 2024.

[6] X.V. Nguyen et al., Effective global approaches for mutual information based feature selection, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **KDD'14**, 512-521, 2014.

[7] H. Venkateswara et al., Efficient approximate solutions to mutual information based global feature selection, *IEEE International Conference on Data Mining*, 1009–1014, 2015.

[8] H. Liu and G. Ditzler, A fast information-theoretic approximation of joint mutual information feature selection, *International Joint Conference on Neural Networks (IJCNN)*, 4610–4617, 2017.

[9] A. B. Finnila et al., Quantum annealing: A new method for minimizing multidimensional functions, *Chemical physics letters*, **219**, 5-6, 1994.

[10] A. Sturm, B. Mummaneni, L. Rullkötter, Unlocking Quantum Optimization: A Use Case Study on NISQ Systems, *arXiv preprint*, arXiv:2404.07171, 2024.