Team shm2024 on Quantum Feature Selection

Notebook for the QuantumCLEF Lab at CLEF 2024

Shimi Gersome[1], Jerin Mahibha[1] and Durairaj Thenmozhi[1]

[1]*Madras Christian College, Chennai,India*

[1]*Meenakshi Sundararajan Engineering College, Chennai, India*

[1]*Sri Sivasubramaniya Nadar college of Engineering, Chennai,India*

Abstract

Quantum has always gained considerable attention in scientific studies as it defies common perception. Quantum Computing- recently evolving- has also started to gain considerable attention. Quantum Computers can solve, some of the distinct challenges in Nondeterministic polynomials (NP)–Hard problems faster than traditional computers. This work is based on the implementation of Task1-feature selection in a shared task, QCLEF2024 with MQ2007 dataset containing 46 features. The task is performed using simulated annealing and Quantum Annealing. The performance of both annealing methods is analyzed based on ndcg@10 (Normalized Discounted Cumulative Gain) and Annealing Time. We obtained a result of 0.3621 and 27222 milliseconds for ndcg@10 and Annealing Time respectively while using Quantum Annealing. We obtained a result of 0.4024 and 284106 milliseconds for ndcg@10 and Annealing Time respectively while using simulated Annealing.

Keywords

Quantum Computing, NP-Hard, Annealing, Simulated, feature selection,

1. Introduction

Quantum is a multidisciplinary research topic revolving around the current era. Presently Quantum computing is an emerging field in Computer Science, which works on the concept of quantum mechanics, it is a combination of Physics, Mathematics, and Computer Science. Quantum computing works on the Quantum bits (qubits), which is the basic unit of Quantum information. The performance of Quantum Computing excels more than typical computers since the computation is done in qubits not in Bits(Binary Digits)[1]. The single qubit holds the state 0,1 or any quantum superposition (through the principle of superposition) of the states and has the potential to compute multiple states concurrently. Quantum Computing algorithms operate on the concept of quantum mechanics superposition and entanglement, which enhance the computational speed exponentially in some distinct areas. Machine Learning(ML) algorithms are intended for predictions, feature selection, classification, and clustering using mathematical formulations. The performance of the ML algorithms is based on data and data size. Quantum Machine Learning(QML) incorporates the ML and Quantum Computing concepts to enable exponentially faster computation. QML aims to optimize current ML techniques by leveraging the performance benefits of quantum algorithms[2]. The shared task of QCLEF2024 Task1 was aimed to perform feature selection by Quantum Computing using MQ2007 Dataset containing 47 features.

Feature selection is an essential process in ML to select the subset of relevant features to improve the model performance from the available features. It reduces the dimensionality, complexity, overfitting,

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France

gshimi2022@gmail.com (S. Gersome); jerinmahibha@gmail.com (J. Mahibha); theni_d@ssn.edu.in (D. Thenmozhi)

and computational cost, which improves model generalization. Similarly QML feature selection decreases the curse of dimensionality, computational complexity, and effect of overfitting and solved through heuristics [3].

2. Related Work

Wang [3] had put forward the idea of Quantum Support Vector Feature Selection(QSVMF) which had optimized maximizing classification accuracy, minimizing selected features and quantum circuit costs, and reducing feature covariance. QSVMF approach had been implemented on the breast cancer dataset and relevant biomarkers had been obtained.

Poggiali et al. [4] had focused on novel approach to Hybrid Quantum Feature Selection(HQFS), done on two synthesized datasets and one real dataset by estimating variance. The ranking produced had been similar to classical algorithms and the low variance features had been perfectly eliminated by HQFS if the number of additional qubits were correctly selected.

Hellstern et al. [5] had recommended a quadratic unconstrained optimization problem (QUBO) for feature selection. For smaller datasets, QUBO had outperformed the classical numerical method. For the Larger dataset Classical method had performed better.

Bhagawati and Subramanian [6] had used Quantum Annealing (QA) to solve the Quadratic Unconstrained Binary Optimization(QUBO), which had adopted optimization theory to minimize or maximize the quadratic objective function. It had been hypothesized that when this hybrid model was used the Feature Selection(FS) had worked comparatively better during the ranking process when using the LETOR dataset. They had found that compared to the standard algorithms(LTR and LamdaMART), the hybrid had yielded a normalized discounted cumulative gain of 0.39 and 0.80, respectively.

3. Methodology

This experimental research is carried out in the workspace provided by the QCLEF2024 using the MQ2007 Dataset for Task1A.[7] LETOR(Learning to Rank for Information Retrieval), is a benchmark dataset on information retrieval for research containing standard features, relevance judgments, data partitioning, evaluation tools, and several baselines. MQ2007 is one of the query sets from LETOR released in 2007 [1].

The MQ2007 dataset contains 47 features among those features one feature is the target feature with the size 41955. The label named target of the MQ2007 dataset is labeled by human experts as “relevant”, “partially relevant”, and “not relevant”. MQ2007 dataset is in text format, and all the features are of datatype float64 except target, which is of int64 datatype containing values 0,1 or 2.

The experimental research of shared Task 1A, feature selection of QCLEF2024 was carried out by Simulated Annealing(SA) and Quantum Annealing(QA), and the performance comparison was made using different parameters and selected the most relevant 5 features from the available features.

3.1. Model Description

The Binary Quadratic Model(BQM) is a collection of binary-valued variables that can hold any two values associated with linear and quadratic biases. Quadratic Unconstrained Binary Optimization (QUBO) models are used by samplers such as the D-Wave system. D-wave uses the process of Quantum Annealing to discover results to solve some distinct types of NP-Hard problems. The problem can be minimized using the minimization formula

[1] https://www.microsoft.com/en-us/research/project/letor-learning-rank-information-retrieval/

$\sum_{i}^{N} q_i x_i + \sum_{i<j}^{N} q_{i,j} x_i x_j$

q_i and $q_{i,j}$ denotes linear and quadratic coefficients.

QUBO is the minimization energy. To extract useful information from the dataset, mutual information and conditional information are used, The dataset is quantified using the Shannon Entropy formula

H(X)=$-\sum_{x \varepsilon X}$ p(x) $log\ p(x)$

p(x) denotes the probability of occurrence of an event.

For Conditional Shannon Entropy (CSE) is calculated if Y is known and X is unknown.

H(X|Y) = H(X,Y) - H(Y)

H(X|Y)=$-\sum_{x \varepsilon X}$ p(x,y) $log\ p(x, y)$ - $H(Y)$

P(x,y) denotes the probability of x and y.

H(X, Y) denotes information X and Y together.

. The features interact with each other using qubits. So the features are quantified using Shannon Entropy(SE) and normalized using Conditional Shannon Entropy(CSE). The correlation or dependency of the target features with the other features is calculated using the mutual information formula

I(X;Y)=$\sum_{y \varepsilon Y} \sum_{x \varepsilon X}$ p(x,y) $log \frac{p(x,y)}{p(x)p(y)}$

p(x) and p(y) denotes marginal probability

p(x,y) denotes joint probability

The variable of interest between the target feature and other features is calculated and stored as another feature. This calculation is performed by the formula

I(X;Y|Z)=$H(X|Z)$ - $H(X|Y, Z)$

H(X|Z) denotes CSE of X on Z

H(X|Y,Z) denotes CSE of X on Y and Z

4. Experimental Setup

4.1. Dataset

The MQ2007, a LETOR (LEarning TO Rank) document retrieval dataset consists of user queries and the corresponding retrieved documents fetched by the query[8]. The retrieved documents are labeled by human experts, as "relevant", "partially relevant", or "not relevant". [9]. The dataset for shared task Task1A, Feature Selection by QCLEF2024 is read from the text file[10].

5. Results

This Task1A is accomplished in the workspace provided by the QCLEF2024 using the MQ2007 Dataset. The dataset has 47 features, among them one is used as a target. We selected only 5 important features

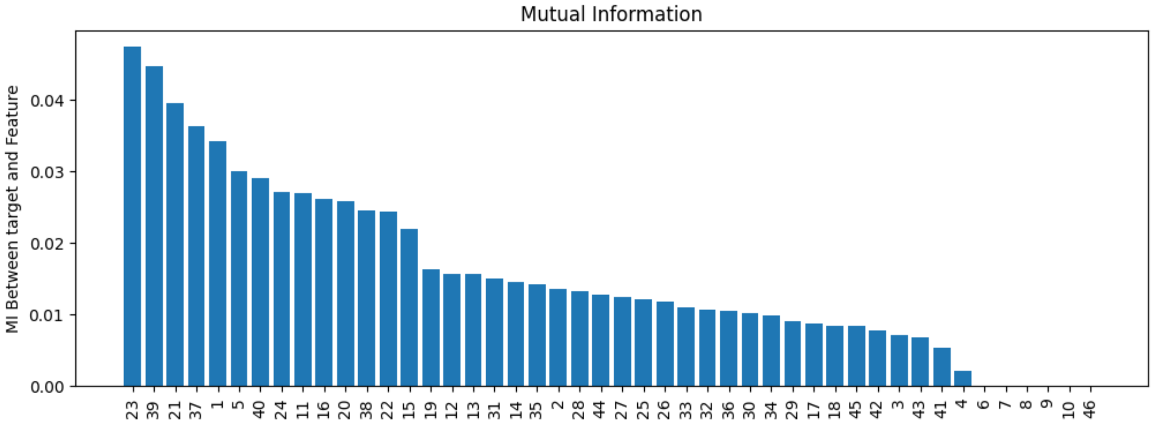

Figure 1: Mutual Information between Target and Feature

Table 1
Performance metrics for Quantum Annealing

Run	ndcg@10	Annealing Time	No. of Features
Run1	0.365	29600	5
Run2	0.3621	27222	5
Run3	0.391	29414	5
Run4	0.3477	28254	5
Run5	0.3245	29315	5

Table 2
Performance metrics for Simulated Annealing

Run	ndcg@10	Annealing Time	No. of Features
Run1	0.4024	224106	5
Run2	0.3032	163964	5
Run3	0.4249	143983	5
Run4	0.4248	146676	5
Run5	0.4205	144050	5

from the 46 features, which are more relevant for future problem solving. The performance metrics of Quantum Annealing are shown in Table 1 and Simulated Annealing is shown in Table 2. The result shows the performance of Simulated Annealing is better when compared with Quantum Annealing in terms of ndcg@10 and Annealing Time. The mutual information between the feature and target is visualized according to rank in Figure 1.

6. Conclusions and Future Work

The shared task, Task 1A by QCLEF2024 intended to perform feature selection from MQ2007_train dataset. The most prominent 5 features are selected from the available 46 features. This implementation is carried out by simulated Annealing and Quantum Annealing and we observed that the annealing time of the Quantum model is less when compared to Simulated Annealing.

References

[1] V. R, H. V, Quantum computing in machine learning - the future of quantum computing, International Journal of Advanced Research in Science, Communication and Technology (2024) 311–314. doi:10.48175/IJARSCT-15955.

[2] M. Nivelkar, S. G. Bhirud, Optimized machine learning: Training and classification performance using quantum computing, in: 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 2021, pp. 8–13. doi:10.1109/ICCCA52192.2021.9666429.

[3] H. Wang, A novel feature selection method based on quantum support vector machine, 2023. arXiv:2311.17646.

[4] A. Poggiali, A. Bernasconi, A. Berti, G. Del corso, R. Guidotti, Quantum feature selection with variance estimation, 2023, pp. 245–250. doi:10.14428/esann/2023.ES2023-99.

[5] G. Hellstern, V. Dehn, M. Zaefferer, Quantum computer based feature selection in machine learning, IET Quantum Communication n/a (????). URL: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/qtc2.12086. doi:https://doi.org/10.1049/qtc2.12086. arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/qtc2.12086.

[6] R. Bhagawati, T. Subramanian, An approach of a quantum-inspired document ranking algorithm by using feature selection methodology, International Journal of Information Technology 15 (2023) 4041–4053.

[7] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, QuantumCLEF 2024: Overview of the Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024, 2024.

[8] T. Qin, T.-Y. Liu, Introducing letor 4.0 datasets, 2013. arXiv:1306.2597.

[9] T. Qin, T.-Y. Liu, J. Xu, H. Li, Letor: A benchmark collection for research on learning to rank for information retrieval, Information Retrieval 13 (2010) 346–374.

[10] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, Overview of QuantumCLEF 2024: The Quantum Computing Challenge for Information Retrieval and Recommender Systems at CLEF, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, 2024.