NICA at Quantum Computing CLEF Tasks 2024

Notebook for the NICA team at Quantum Computing Lab CLEF 2024

Aylin Naebzadeh[1,*], Sauleh Eetemadi[2]

[1]Student at School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.
[2]Assistant Professor of Computer Science, School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.

Abstract

We present the models implemented by the NICA group for the Quantum Computing (QuantumCLEF) Shared Task at CLEF 2024. Our participation focused on Task 1A: Feature Selection (Information Retrieval Task). We propose a feature selection algorithm based on a quadratic unconstrained binary optimization (QUBO) problem, which selects a specified number of features considering their importance and redundancy. This task was solved using a real quantum computer provided by D-Wave on the MQ2007 and ISTELLA datasets. Our approach utilized "Shannon Entropy" to target the mutual information between each feature and the target value. In QuantumCLEF Task 1A, the organizers suggested training a LambdaMART model on the selected features and evaluating performance using the nDCG@10 metric. Our team achieved nDCG@10 scores of 0.4506 and 0.6211 for the MQ2007 and ISTELLA datasets, respectively.

Keywords

Quantum Computing, Feature Selection, Information Retrieval, QUBO, Quantum Annealer, D-Wave, Shannon Entropy, nDCG@10

1. Introduction

Machine learning (ML) models are highly effective in various data analytics tasks, including classification, regression, and data generation. However, the performance and resource requirements of these models often scale with the number of input features. Models with a larger number of features demand more memory and computational power during training. In applications with strict resource limitations, such as embedded systems, it is crucial to develop small and efficient models. Consequently, a common objective in ML pipelines is to reduce the number of features while minimizing information loss through a process known as dimensionality reduction [1, 2].

An important example for such a strategy is feature selection (FS), where the input dimension is reduced by selecting only a subset of all available features without performing additional transformations [3]. Despite its importance, feature selection poses significant challenges for classical computers. One major limitation is that feature selection is an NP-hard problem, meaning that the computational effort required to find the optimal subset of features grows exponentially with the number of features. As a result, solving this problem can be extremely time-consuming, especially for large datasets with high-dimensional feature spaces. Additionally, traditional algorithms often require substantial computational resources and memory, making them impractical for real-time applications or systems with limited resources. These challenges underscore the need for more efficient and scalable approaches to feature selection.

To address this issue, a new challenge called QuantumCLEF [1] [4, 5, 6] has been established with the objective of efficiently and effectively employ quantum annealers for problems like features selection in *Information Retrieval (IR)* tasks and *Recommender Systems (RS)*.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
a_naeb@comp.iust.ac.ir (A. Naebzadeh); sauleh@iust.ac.ir (S. Eetemadi)
https://github.com/AylinNaebzadeh (A. Naebzadeh); http://www.sauleh.ir/ (S. Eetemadi)

[1]https://qclef.dei.unipd.it

Quantum Computing (QC) has garnered significant attention from researchers across various fields, as technological advancements have made QC resources more accessible and applicable to practical problems. In the current landscape, IR and RS require computationally intensive operations on massive and heterogeneous datasets. Consequently, Quantum Computing, particularly *Quantum Annealing (QA)* technologies, holds the potential to enhance these systems' performance in terms of both efficiency and effectiveness [4].

This paper describes the NICA team's contribution to the first shared task of QuantumCLEF 2024. The objective of this task is formulating the well-known NP-Hard feature selection problem to solve it with a quantum annealer and compare the results with a simulate annealing. Two datasets have been defined for the first part of this task, MQ2007 [7] and ISTELLA Letor [8]. These datasets contain pre-computed features and the objective is to select a subset of these features to train a learning model, such as LambdaMART [9] or a content-based RS, and to achieve best performance according to metrics such as nDCG@10. Under the given shared task, our system selects the most relevant features based on their mutual information with the target value, and then submits those features to a D-Wave quantum computer[2]. Although D-Wave is not freely available, the organizers of this challenge made it accessible to us for the purpose of this competition.

This work is structured as follows: Section 2 briefly provides a description of several earlier studies. Section 3 will then present an explanation of task description. Following that, Section 4 and 5 will outline the experimental methodology and evaluation results respectively. Finally, in Section 6, we will present the key findings and conclusions of our studies, as well as some potential directions for future research.

2. Related Works

Quantum computing has been widely applied across various domains such as energy applications [10] and finance [11] . In the field of computer science, quantum computing has been explored for applications like *Support Vector Machine (SVM)* as machine learning algorithms [12, 13]. However, there is a limited amount of research on the applications of QC and QA in the domains of IR and RS [14, 15, 16, 17].

In a study by Nembrini et al. [14], QA was applied to IR and RS tasks such as feature selection, demonstrating the feasibility and promising improvements in efficiency and effectiveness. Additionally, Mücke et al. [2] conducted a series of numerical experiments using classical computers, quantum gate computers, and quantum annealers. Their results showed competitive performance when comparing several standard methods on various benchmark datasets, highlighting the potential of quantum technologies in enhancing IR and RS tasks.

3. Tasks Description

This is the first edition of QuantumCLEF. In this lab, there are two tasks involving computationally intensive problems closely related to the Information Access field: **Feature Selection** and **Clustering**. Each task presents a problem that can be solved using the QA paradigm. Participants are required to submit their solutions using both QA and Simulated Annealing (SA) to compare the two methods in terms of efficiency and effectiveness. Figure 1 provides an overview of the tasks and datasets in QuantumCLEF 2024. More detailed descriptions of the tasks are available on the organizers' QuantumCLEF website[3] and in this paper by Pasin et al. [4].

[2]https://www.dwavesys.com/solutions-and-products/cloud-platform
[3]https://qclef.dei.unipd.it/index.html/

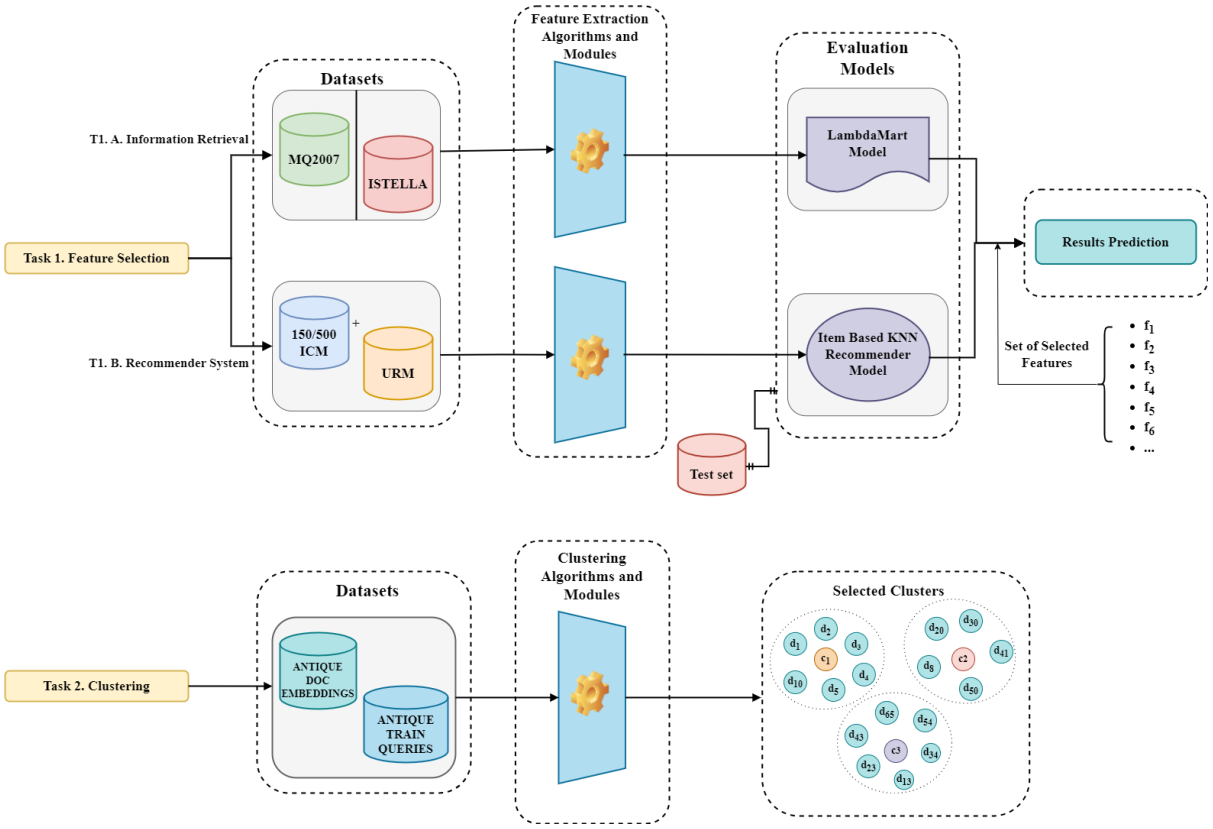


Figure 1: An Overview of Tasks in QuantumCLEF 2024

Table 1
MQ2007 Dataset

	class	1	2	3	4	...	43	44	45	46
0	0	0.000000	0.0	0.0	0.0	...	0.055556	0.000000	0.00	0.0
1	1	0.205882	0.0	0.0	0.0	...	0.000000	0.333333	0.42	0.0
2	0	0.000000	0.0	0.0	0.0	...	0.092593	0.500000	0.68	0.0
3	1	0.088235	0.0	0.0	0.0	...	0.000000	0.666667	0.86	0.0
4	0	0.558824	0.0	0.0	0.0	...	0.000000	0.500000	0.46	0.0

Table 2
ISTELLA Dataset

	target	1	2	3	4	...	217	218	219	220
0	0	1.797693e+308	1008.0	2.0	288.0	...	1.0	0.000000	0.0	0.0
1	0	1.797693e+308	105.0	2.0	0.0	...	1.0	0.333333	0.0	0.0
2	0	1.797693e+308	1058.0	1.0	5248.0	...	1.0	0.500000	0.0	2.0
3	0	1.797693e+308	121.0	0.0	0.0	...	1.0	0.666667	0.0	0.0
4	0	1.797693e+308	127.0	0.0	0.0	...	1.0	0.500000	0.0	0.0

4. Experimental Setup

4.1. Task 1.A. Dataset

4.1.1. MQ2007

The MQ2007 dataset comprises 46 independent features and a binary dependent feature labeled “class”. Table 1 provides example values within this dataset.

Table 3
QUBO Representation for Feature Selection

	Optimization	Linear Terms	Quadratic Terms	Formula
Feature Selection	Maximize	$I(X_i;Y)$	$I(X_j;Y\|\|X_i)$	$\sum_{i\in S}\left\{I(X_i;Y)+\sum_{j\in S,j\neq i}I(X_j;Y\|\|X_i)\right\}$
QUBO	Minimize	$q_i x_i$	$q_{i,j}x_i x_j$	$\sum_i^N q_i x_i+\sum_{i<j}^N q_{i,j}x_i x_j$

4.1.2. ISTELLA

The ISTELLA dataset consists of 220 independent features and a binary dependent feature labeled "target". Table 2 presents example values within this dataset. This dataset poses an additional challenge due to the large number of features, which cannot be directly accommodated in the Quantum Processing Unit (QPU). To mitigate this issue, we applied a filtering process to select only the most relevant features. Specifically, We extracted 50 features based on their mutual information with the "target" feature, ensuring a threshold of 0.031, and we used only the selected features for subsequent analysis. The threshold of 0.031 was carefully chosen after testing various thresholds to find a balance between retaining a sufficient number of informative features and managing the limitations of our QPU. By setting the threshold to 0.031, we ensured that we retained enough features to maintain the integrity and predictive power of our model. At the same time, we avoided the pitfall of including too many features, which would exceed the capacity of the QPU and hinder our ability to perform efficient quantum computations.

4.2. Workspace

In this challenge every team participating was provided with an identical workspace, ensuring equal access to resources. Recognizing that hardware resource limitations are a common constraint in research, organizers established a custom infrastructure to mitigate this issue. This infrastructure was necessary because participants could not have direct access to quantum annealers, and organizers aimed to ensure that measurements were fair and reproducible.

Each team's workspace, accessible via a browser with the correct credentials, included a preconfigured Git repository essential for maintaining reproducibility. A centralized dispatcher managed and tracked all team submissions, using a secret API key for submitting problems to the quantum annealer, ensuring participants remained unaware of this key. Additionally, a web application served as the primary information hub, allowing teams to view their quotas and various statistics through a dashboard. Organizers had their own dashboard to manage teams and tasks effectively [18]. Figure 2 provides a high-level representation of the infrastructure.

5. Methodology

5.1. Quantum Annealing

QA is a quantum computing paradigm that relies on specialized devices known as quantum annealers, designed to address optimization problems. In QA, a problem is encoded as the energy landscape of a physical system, and quantum-mechanical principles are employed to guide the system towards a state of minimal energy, which corresponds to the solution of the original problem. To utilize quantum annealers effectively, problems must be formulated as minimization tasks using the *Quadratic Unconstrained Binary Optimization (QUBO)* formulation, which is defined as follows:

$$\min y = x^T Q x \tag{1}$$

where x is a vector of binary decision variables and Q is a matrix of constant values representing the problem to solve [4].

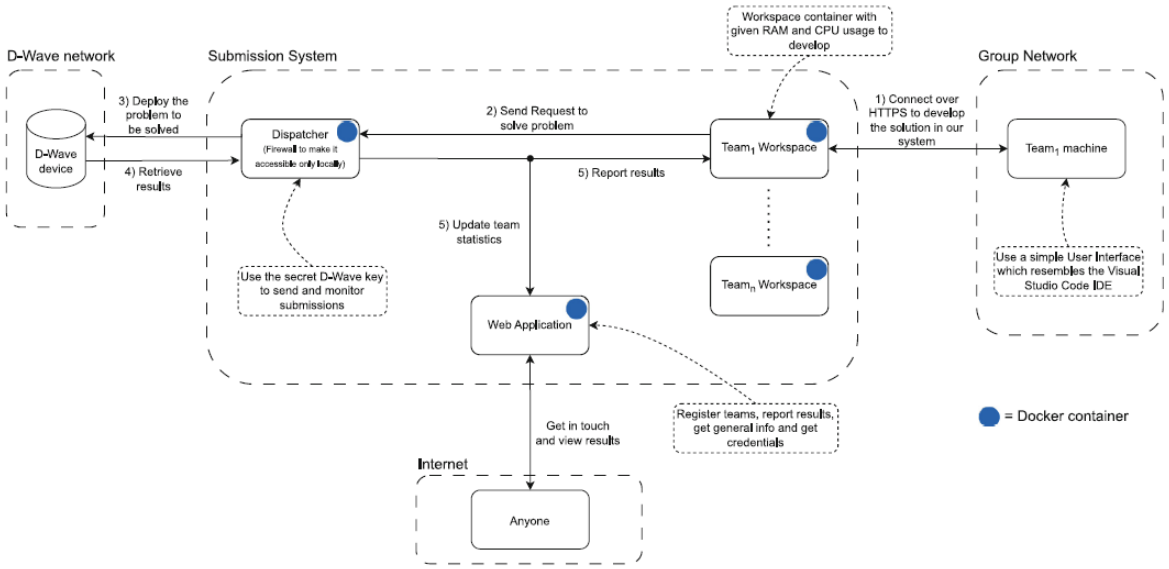


Figure 2: High-level representation of the infrastructure [18].

5.2. QUBO Representation for Feature Selection

D-Wave systems solve Binary Quadratic Models (BQM). Given N variables $x_1, ..., x_N$, where each variable x_i can have binary values 0 or 1, the system finds assignments of values that minimize,

$$\sum_{i}^{N} q_i x_i + \sum_{i<j}^{N} q_{i,j} x_i x_j \tag{2}$$

where q_i and $q_{i,j}$ are configurable (linear and quadratic) coefficients. To formulate a problem for the D-Wave system is to program q_i and $q_{i,j}$ so that assignments of $x_1, ..., x_N$ also represent solutions to the problem. For feature selection, the Mutual Information QUBO (MIQUBO) method formulates a QUBO based on the approximation above for $I(X_k; Y)$, which can be submitted to the D-Wave quantum computer for solution. The reduction of scope to permutations of three variables in this approximate formulation for MI-based optimal feature selection makes it a natural fit for reformulation as a QUBO. Table 3 shows a QUBO representation for feature selection task.

5.3. Computing Mutual Information

We establish and employ several functions to calculate the values of Mutual Information and Conditional Mutual Information, which are integral to defining our problem, and have been given by the organizers in the workspace. Specifically, we outline some helpful definitions below.

5.3.1. Quantifying Information: Shannon Entropy

Shannon entropy[4] [19], denoted as $H(X)$, is a mathematical measure that quantifies the information content of a signal:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3}$$

Here, $p(x)$ represents the probability of an event's occurrence. The Shannon Entropy formula can be interpreted as assigning a value of $\log \frac{1}{p(x)}$ to each event based on its probability, weighted by that

[4] https://en.wiktionary.org/wiki/Shannon_entropy

probability. The reciprocal in the logarithm ensures that less likely events are attributed with more information.

5.3.2. Conditional Shannon Entropy

Conditional Shannon Entropy (CSE) quantifies the information content of one signal, X, given the value of another signal, Y:

$$H(X|Y) = H(X,Y) - H(Y) \quad = -\sum_{x \in X} p(x,y) \log p(x,y) - H(Y) \tag{4}$$

Here, the joint Shannon Entropy, $H(X,Y)$, represents the combined information content of both signals, where $p(x,y)$ denotes their joint probability. For instance, knowing that it is winter reduces the informational value of news regarding rainfall.

5.3.3. Mutual Information

Mutual information [5] [20] between variables X and Y is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5}$$

Here, $p(x)$ and $p(y)$ represent the marginal probabilities of X and Y, respectively, while $p(x,y)$ denotes the joint probability.

Alternatively, mutual information can be expressed as:

$$I(X;Y) = H(Y) - H(Y|X) \tag{6}$$

In this equation, $H(Y)$ represents the Shannon Entropy of Y, and $H(Y|X)$ denotes the Conditional Shannon Entropy of Y given X.

Mutual information quantifies the amount of information about one random variable that can be inferred from observations of another. Intuitively, when the mutual information between two variables is high, a model based on either variable alone can effectively capture their combined contribution. High mutual information implies a strong dependence or correlation between the variables, indicating that knowledge of one variable provides significant information about the other. Conversely, a low or zero mutual information value suggests little to no association between the variables.

5.3.4. Conditional Mutual Information

Conditional Mutual Information (CMI) between a variable of interest, X, and a feature, Y, given the selection of another feature, Z, is calculated as:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{7}$$

Here, $H(X|Z)$ represents the Conditional Shannon Entropy (CSE) of X given Z, while $H(X|Y,Z)$ denotes the CSE of X conditional on both Y and Z.

A high conditional mutual information value indicates strong dependence or correlation between variables X and Y given the value of variable Z. It suggests that knowledge of variable Z provides significant information about the relationship between variables X and Y. Conversely, a low or zero conditional mutual information value suggests little to no association between variables X and Y given the value of variable Z.

[5] https://en.wikipedia.org/wiki/Mutual_information

Table 4
Results for Task 1. A. Information Retrieval

Dataset	Submission ID	nDCG@10	Annealing Time (us)	Type	# Features
MQ2007	6f7d7d44-c559-4e36-9b10-b7e51e521036	0.4506	274119	QA	17
MQ2007	169	0.4498	3509658	SA	15
ISTELLA	c5888bf1-4549-418c-92b8-b7175c9185e4	0.596	427442	QA	15
ISTELLA	380	0.6211	3998441	SA	15

6. Results

The overview of the final results of this study's submissions can be found in the Table 4. The Annealing time measures the execution time of the approach. In the case of QA this consists in the programming time, sampling time and post-processing time. The results of our study indicate that QA outperforms SA in terms of both effectiveness and efficiency. QA demonstrated superior capability in finding optimal solutions more consistently and quickly, showcasing its potential for solving complex optimization problems more effectively. The efficiency gains were evident through faster convergence times and reduced computational resources required compared to SA. This highlights QA as a more powerful and efficient approach for tackling intricate problems within the context of our experiments.

7. Conclusion and Future Work

Feature selection plays a pivotal role in machine learning, significantly impacting the performance of models by reducing overfitting, reducing the effect of curse of dimensionality, and reducing computational complexity. The task, however, is NP-Hard, necessitating efficient and innovative solutions. The emerging field of quantum computing offers promising avenues for tackling this challenge. In our study, we leveraged Shannon entropy to construct our Binary Quadratic Model (BQM) object, a novel approach in the realm of feature selection. We then employed two distinct methods to solve our problem: SA and QA, provided by D-Wave. Our results underscore the potential of QA in providing superior solutions for complex problems like feature selection. As we look to the future, we anticipate further exploration and refinement of quantum methods for feature selection, contributing to the advancement of machine learning and, by extension, numerous fields that depend on it.

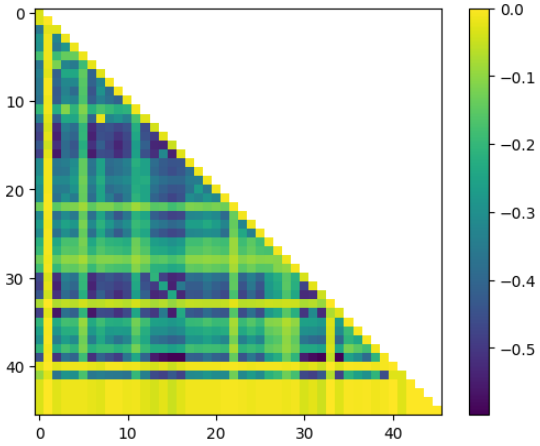
Acknowledgments

We would like to express our deep appreciation to the organizers of QuantumCLEF 2024 for providing us with the opportunity to participate in this esteemed event, and sharing informative tutorials, documentations, and codes with us. It was all a great and valuable experience to work with the provided cloud based infrastructure and quantum computers, and gain new knowledge. We are truly grateful for the chance to present our work and contribute to the progress of our field. We would like to once again acknowledge and appreciate the support and encouragement received from all individuals involved, as their contributions have been instrumental in our accomplishments.

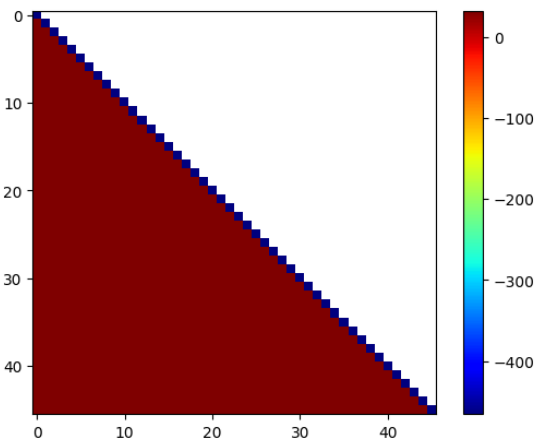
References

[1] L. Van Der Maaten, E. Postma, J. Van den Herik, et al., Dimensionality reduction: a comparative, J Mach Learn Res 10 (2009).
[2] S. Mücke, R. Heese, S. Müller, M. Wolter, N. Piatkowski, Feature selection on quantum computers, Quantum Machine Intelligence 5 (2023) 11.

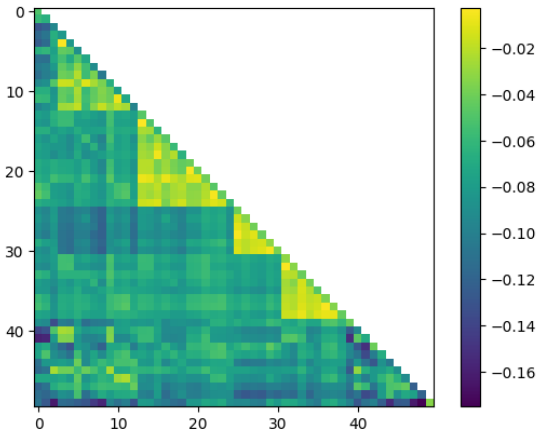
[3] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Computers & electrical engineering 40 (2014) 16–28.
[4] A. Pasin, M. Ferrari Dacrema, P. Cremonesi, N. Ferro, qclef: A proposal to evaluate quantum annealing for information retrieval and recommender systems, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 97–108.
[5] 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.
[6] 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.
[7] T. Qin, T.-Y. Liu, Introducing letor 4.0 datasets, arXiv preprint arXiv:1306.2597 (2013).
[8] D. Dato, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, R. Venturini, Fast ranking with additive ensembles of oblivious and non-oblivious regression trees, ACM Transactions on Information Systems (TOIS) 35 (2016) 1–31.
[9] C. J. Burges, From ranknet to lambdarank to lambdamart: An overview, Learning 11 (2010) 81.
[10] H. P. Paudel, M. Syamlal, S. E. Crawford, Y.-L. Lee, R. A. Shugayev, P. Lu, P. R. Ohodnicki, D. Mollot, Y. Duan, Quantum computing and simulations for energy applications: Review and perspective, ACS Engineering Au 2 (2022) 151–196.
[11] R. Orús, S. Mugel, E. Lizaso, Quantum computing for finance: Overview and prospects, Reviews in Physics 4 (2019) 100028.
[12] D. Willsch, M. Willsch, H. De Raedt, K. Michielsen, Support vector machines on the d-wave quantum annealer, Computer physics communications 248 (2020) 107006.
[13] A. Delilbasic, B. Le Saux, M. Riedel, K. Michielsen, G. Cavallaro, A single-step multiclass svm based on quantum annealing for remote sensing data classification, IEEE journal of selected topics in applied earth observations and remote sensing (2023).
[14] R. Nembrini, M. Ferrari Dacrema, P. Cremonesi, Feature selection for recommender systems with quantum computing, Entropy 23 (2021) 970.
[15] M. F. Dacrema, A. Pasin, P. Cremonesi, N. Ferro, Using and evaluating quantum computing for information retrieval and recommender systems (2024).
[16] R. Pellini, M. F. Dacrema, P. Cremonesi, Towards improved qubo formulations of ir tasks for quantum annealers (2023).
[17] M. Melucci, et al., Introduction to information retrieval and quantum mechanics, Springer, 2015.
[18] A. Pasin, M. F. Dacrema, P. Cremonesi, N. Ferro, Quantumclef-quantum computing at clef, in: European Conference on Information Retrieval, Springer, 2024, pp. 482–489.
[19] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1991) 145–151.
[20] T. E. Duncan, On the calculation of mutual information, SIAM Journal on Applied Mathematics 19 (1970) 215–220.



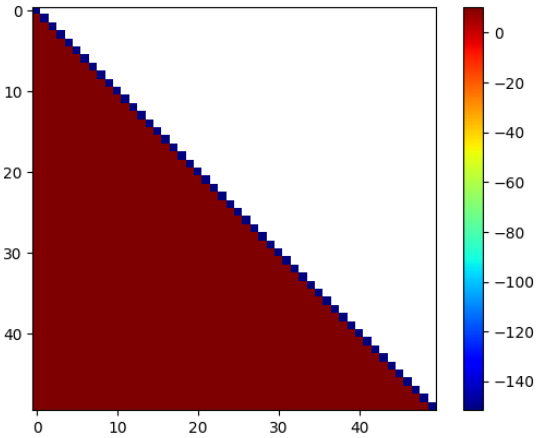
(a) Matrix for MQ2007 with All Features.



(b) Matrix for MQ2007 to make our model choose 15 Features.

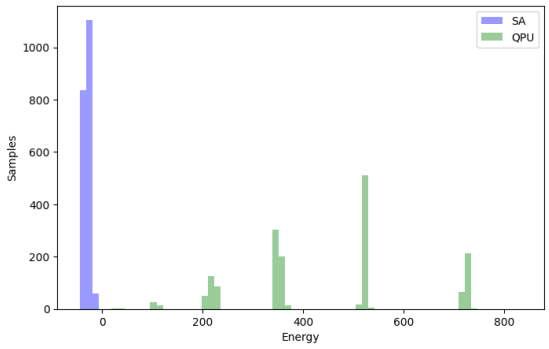


(c) Matrix for ISTELLA with All Selected 50 Features.

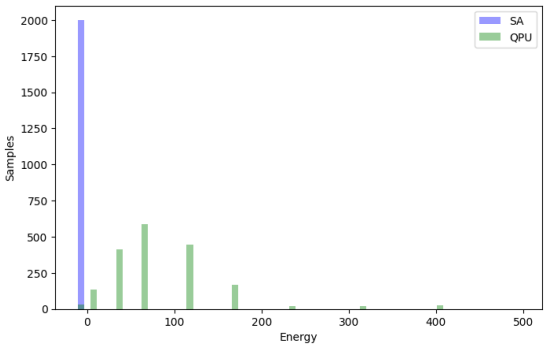


(d) Matrix for ISTELLA to make our model choose 15 Features.

Figure 3: QUBO Matrices



(a) Energy Plot for SA and QPU Samples Over MQ2007



(b) Energy Plot for SA and QPU Samples Over ISTELLA

Figure 4: Overview of Comparison between Computed Simulated Annealing (SA) and Quantum Annealing (QA) Samples Using Histogram Energy Reporting Levels