QUANTUM COMPUTING INC

Fraud Detection using QCi's Dirac-3

July 2024

Quantum Computing Inc
quantumcomputinginc.com
(703) 436-2161

1 Introduction

The QBoost algorithm, introduced by Neven et al. (2009), is a classification method that leverages quadratic optimization solvers (such as QCi's Dirac machines) to obtain superior speed and power consumption over classical approaches. QBoost is an adaptation of the classical boosting algorithm. Boosting is a powerful technique in machine learning that combines the output of several weak classifiers to produce a strong classifier. In classical boosting algorithms, such as AdaBoost, the weights of the weak classifiers are adjusted iteratively based on their performance, with the goal of minimizing the overall classification error.

The innovation of QBoost lies in its utilization of quantum computing to perform this optimization. By encoding the boosting problem into a quadratic optimization problem, QBoost exploits the Dirac machine's ability to explore multiple solutions simultaneously and escape local minima more efficiently than classical algorithms.

In what follows, the Dirac-3 implementation of QBoost is discussed and used to perform a binary classification on a credit card fraud dataset.

2 Formulation

The idea is based on the concept of boosting. Let us assume that we have a collection of N "weak" classifiers h_i where $i = 1, 2, ..., N$. The goal is to construct a "strong" classifier as a linear superposition of these weak classifiers, that is,

$$y = \sum_{i=1}^{N} w_i h_i(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is a vector of input features and $y \in \{-1, 1\}$. The goal is find w_i, weights associated with the weak classifiers.

We use a training set $\{(\mathbf{x_s}, y_s)|s = 1, 2, ..., S\}$ of size S. We can determine optimal weights w_i by minimizing,

$$\min_{\mathbf{w}} \sum_{s=1}^{S} |\sum_{i=1}^{N} w_i h_i(\mathbf{x_s}) - y_s|^2 + \lambda \sum_{i=1}^{N} (w_i)^2 \tag{2}$$

where the regularization term $\lambda \sum_{i=1}^{N} (w_i)^2$ penalizes non-zero weights; λ is the regularization coefficient.

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} w_i w_j + \sum_{i=1}^{N} C_i w_i \tag{3}$$

where

$$J_{ij} = \sum_{s=1}^{S} h_i(\mathbf{x_s}) h_j(\mathbf{x_s}) \tag{4}$$

and

$$C_i = -2 \sum_{s=1}^{S} y_s h_i(\mathbf{x_s}) \tag{5}$$

subject to,

$$\sum_{i=1}^{N} w_i = 1 \tag{6}$$

Note that the above algorithm assumes that the total number of weak classifiers, that is N, is less than the number of available qudits on Dirac-3.

2.1 Choices of Weak Classifiers

There are many ways to design a subset of weak classifiers. We have tested QBoost using logistic regression, decision tree, naive Bayesian, and Gaussian process classifiers. Each weak classifier is constructed using one or two of the features chosen from all features. This yields a set of weak classifiers that can be used to construct a strong classifier.

3 Use Case

3.1 Dataset

The Kaggle Credit Card Fraud Detection dataset is a popular dataset used for machine learning research and practice, particularly in the field of anomaly detection.

The dataset contains transactions made by European credit cardholders over a period of two days in September 2013. It is comprised of a total of 284,807 transactions, with 492 of them (approximately 0.172%) being fraudulent. This severe imbalance makes it a suitable dataset for practicing techniques related to imbalanced classification problems.

The dataset includes 31 columns, which are:

- Time: The time elapsed in seconds from the first transaction in the dataset. This feature helps in understanding the transaction sequence.
- V1 to V28: These are the principal components obtained after applying Principal Component Analysis (PCA) to the original feature set for privacy protection and dimensionality reduction. The exact nature of these features is not provided due to confidentiality issues.
- Amount: The transaction amount, which can be useful for making decisions on whether a transaction is fraudulent or not.
- Class: The target variable, where 1 indicates that the transaction is fraudulent and 0 indicates that it is not.

3.2 Data Imbalance

The dataset is highly imbalanced, with the majority of transactions being non-fraudulent. The primary challenge with this dataset is dealing with the class imbalance while training models. Various techniques can be employed. In the current limited study, we use sub-sampling of the majority class (non-fraud cases) to train the model.

3.3 Benchmarking against Classical Alternatives

It is important to compare the performance of QBoost to a state-of-the-art classical method. Here, we have used XGBoost, a ensemble learning method which is widely used for classification problems.

3.4 Results

We created a dataset by sub-sampling of the majority class. We randomly chose a balanced dataset consisting of about 3000 samples. The 80% of samples were used for training and the remaining was used for testing. Repeating the training using multiple randomly chosen samples yields very similar results. The table below shows accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) on testing data for both QBoost and XGBoost models. The figure shows the ROC curve corresponding to QBoost and XGBoost.

Model	Accuracy	Precision	Recall	F1 Score	AUC
QBoost	0.84	0.96	0.70	0.81	0.87
XGBoost	0.84	0.87	0.78	0.82	0.87

Table 1: Comparison of QBoost and XGBoost results on credit card fraud dataset. All statistics are calculated on the test data.

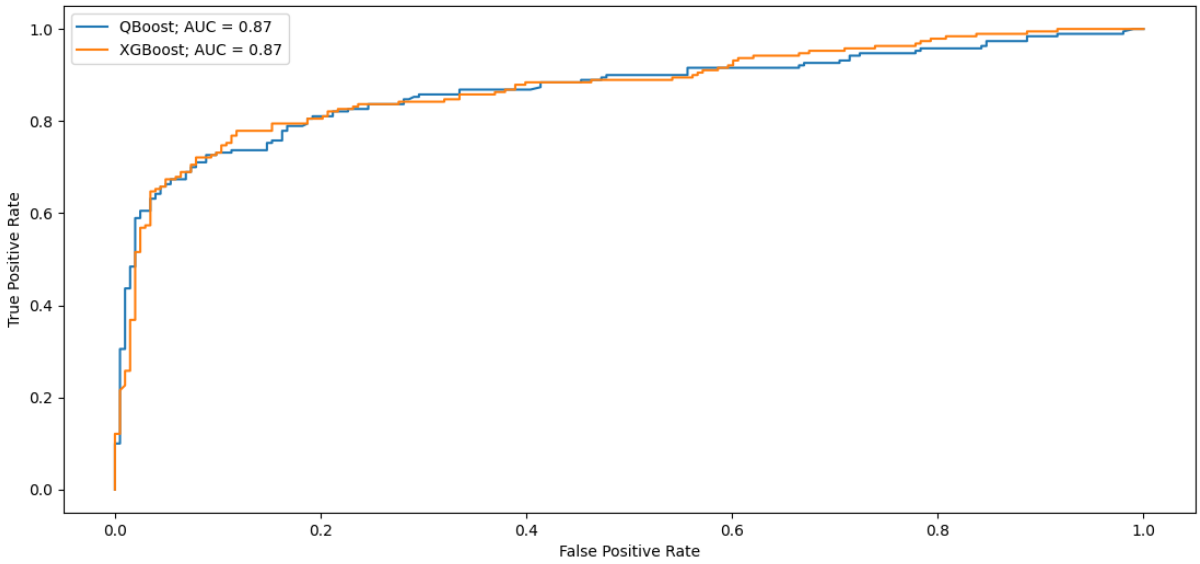

Figure 1: ROC curves for QBoost vs. XGBoost. The curves are computed based on the test data.

3.5 Results

We have presented a QCi's Dirac-3 based classification model. The performance of the model on a highly imbalanced fraud dataset is reasonable and is similar to the performance of the classical XGBoost.