機械学習手法としてのQBoostの活用

理学専攻・情報科学コース　2240649　木内美波

1　はじめに

近年注目されている量子コンピュータは0と1で情報を表す一般的なコンピュータと異なり, 代わりに量子ビットの重ね合わせの状態を用いて状態を表す. 量子コンピュータ技術は大きく分けて二通り存在する. 量子ゲート方式と量子アニーリング方式である.

量子アニーリング方式とは組み合わせ最適化問題をイジングモデルやQUBO (Quadratic Unconstrained Binary Optimization) 形式で表現することで解く手法である. 組み合わせ最適化問題とは膨大な選択肢の中から制約を満たし, コストを最小化する最適な解の組み合わせを求める問題である [1]. 量子アニーリングは量子揺らぎによる状態の変化を利用して組み合わせ最適化の近似解を求めることができる. 組み合わせ最適化問題はあらゆる分野で登場するため, これを解くことができる量子アニーリング技術への期待は極めて高まっている.

量子アニーリングは最適化問題として定式化されることの多い機械学習にも効果的であると言われている. 本研究では量子アニーリングをアンサンブル学習に取り入れたQBoost [2] という機械学習の手法を用いて分類問題に取り組んだ.

2　QBoost

QBoostとは量子アニーリングを用いたアンサンブル学習の一つである. アンサンブル学習とは精度の低い弱学習器をいくつか組み合わせることで精度の高い強学習器を作る手法である. 弱学習器の候補の中から, どの学習器を使用するかは組み合わせ問題であり, 量子アニーリングを用いて, 最適解に近い解を求めることができる. 一部の学習データのみを使用し, 軽い処理で作成した弱学習器を使用するのもQBoostの特徴である. QBoostを回帰問題に応用させたモデルは次式である.

$$C\left(\vec{x}\right)=\sum_{i=1}^{N} w_i c_i\left(\vec{x}\right) \tag{1}$$

N 個の弱学習器の集合を $\{c_i\}\,(i=1,\ldots,N)$ としている. ここで, あるデータ $\vec{x}$ に対しての弱学習器の結果を $c_i\left(\vec{x}\right)$ と表す. そして, 各学習器の重み (採用するかしないか) を $w_i \in \{0,1\}\,(i=1,\ldots,N)$ としている. つまり使用する弱学習器の結果の合計の符号が強学習器 $C\left(\vec{x}\right)$ の結果になる.

弱学習器の組合せ最適化は, 以下のハミルトニアンを最小化することで行う.

$$H\left(\vec{w}\right)=\sum_{d=1}^{D}\left(\frac{1}{N}\sum_{i=1}^{N} w_i c_i\left(\vec{x}^{(d)}\right)-y^{(d)}\right)^2+\lambda\sum_{i}^{N} w_i \tag{2}$$

D 個の学習データの集合を $\{\vec{x}^{(d)}\}\,(d=1,\ldots,D)$, 対応する正解ラベルを $\{y^{(d)}\}\,(d=1,\ldots,D)$ とする. 第一項は強学習器と正解ラベルの差を最小化するための, 第二項は最終的に強学習器に採用する弱学習器の個数を絞りこむための項である. w_i は0か1をとる二値変数であり, 使用する弱学習器の組合せを表す. λ は正の実数である. λ が大きいほど w_i は0をとることが増えるため使用する弱学習器の数が減少する. つまり, 強学習器と正解ラベルの差が少なく, 使用する弱学習器の数が少ないほど良い, というモデルである. 本研究では, QBoostを使用した手法が古典的な方法より精度の高い結果になるかどうかを確認する.

3　二値分類への応用

日本経済新聞が提供する日経電子版の利用状況からユーザの年代を推定する二値分類問題に取り組んだ. QBoostはLightGBMに比べ, シンプルで解釈性に優れていることが特徴である. まず全ての弱学習器を用いた結果をbaseとする. 次に学習し選別を行った弱学習器を学習データに適用した結果をtrain, テストデータに適用させた結果をtestとする. 表1では弱学習器を選別したtestの方が全ての弱学習器を用いたbaseより精度が高いことが分かった. 使用する特徴量を変更したり, ハミルトニアンのパラメタを調整することで, QBoostを用いた分類の精度が向上する可能性がある.

表 1: 予測精度の結果 (Accuracy)

	base	train	test	lightGBM
20代と50代	0.595	0.649	0.646	0.689
30代と50代	0.550	0.618	0.622	0.644

4　回帰分類への応用

4.1　データの準備

使用したのはscikit-learnのオープンソースで, カリフォルニアの住宅価格データセット, ボストンの住宅価格データセット, 糖尿病の進行に関するデータセットの3種類である. カリフォルニアの住宅価格のデータセットは部屋数や築年数などの項目と住宅価格の目的変数からなるデータセットである. ボストンの住宅価格のデータセットは犯罪率や部屋数などの項目と住宅価格の目的変数からなるデータセットである. 糖尿病のデータセットは年齢や血圧などの項目と一年後の糖尿病進行度の目的変数からなるデータセットである. 計算にはシミュレーテッドアニーリングを用いた.

4.2　パラメタ

精度を上げるためにパラメタの調整を行った. 調整が必要なパラメタは次の4つである. 一つ目はハミルトニアンのパラメタ λ である. 二つ目は弱学習器のモデルを作成するために使用するデータのサンプル数である. 三つ目は作成する弱学習器の決定木の深さである. 四つ目は作成する弱学習器の個数である. この弱学習器から更にQBoostでいくつかを選別して弱学習器を作成する. ハミルトニアンの λ と精度の関係を観察したところ, λ と精度との間に明確な相関関係は見

られなかった．サンプル数に関しては，比較した範囲ではサンプル数が多いほど精度がよくなる結果となった．同様に決定木の深さと弱学習器の個数で比較を行った．R2_score と RMSE で値を比較したところ，決定木が深いほど精度がよくなることが分かった．しかし，軽い処理で弱学習器を作成したいという目的のためできるだけ深さは浅い方が良い．

4.3 結果

λ と精度の関係，λ と選ばれる弱学習器の関係を調べるためにグラフを作成した．図 1 は横軸を λ，縦軸を RMSE とした．決定木の深さは 5，弱学習器の個数は 30 個，弱学習器の作成に使用するサンプル数は 1000 個に固定した．まず全ての弱学習器を用いた結果を base とする．次に学習し選別を行った弱学習器を学習データに適用した結果を train，テストデータに適用させた結果を test とする．base より test の方が RMSE が小さければ，QBoost による弱学習器の選別が精度の向上に繋がったと言える．また，test より train の RMSE が大幅に小さくなってしまった場合は過学習を起こしている可能性がある．極端に学習データに対応した学習をしてしまい学習データ以外の一般的なデータに対応できない結果になってしまっているため，この場合は上手く学習したとは言えない．カリフォルニアのデータでは base の精度を test が上回っており，全ての弱学習器を使用するより選別する方が良い精度であることが分かる．図 2 では横軸を λ，縦軸を QBoost で選ばれた弱学習器の数とした．λ が多いほど選ばれる弱学習器の数が少なくなっているのが確認できた．図 1 の右端のデータは選ばれた弱学習器の数が極端に少なくなり高い精度が出なくなったと考えられる．

5 まとめ

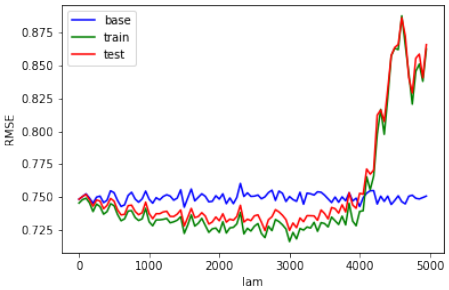
QBoost を二値分類問題と回帰問題に応用し結果を観察した．二値分類問題に応用した実験では，ユーザーの年代推定問題に取り組んだ．弱学習器の選別を行うことで精度の向上に繋がることが確認できた．回帰問題に応用した実験では 3 種類のデータセットを用いた実験を行った．パラメタチューニングが難しかったが，カリフォルニアのデータでは二値分類問題と同様に弱学習器を選別した方が精度が高くなった．今後は違う方法でのパラメタの調整などを行い，ボストンと糖尿病のデータセットでも良い精度が出るようにしたい．

参考文献

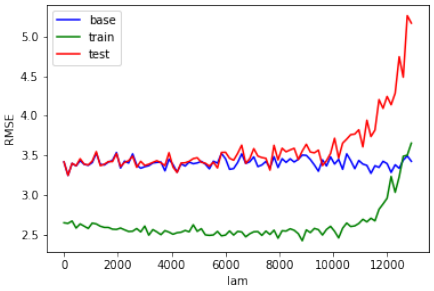
[1] 田中宗，松田佳希，量子アニーリングの動作原理と応用探索，計測と制御 **58**，203 (2019)

[2] H. Neven et al, Qboost: Large scale classifier training withadiabatic quantum optimization, PMLR25, 333 (2012)

(a)



(b)



(c)

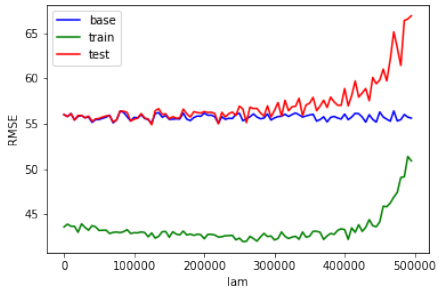


図 1: RMSE の λ 依存性．データセットはそれぞれ (a) カリフォルニア，(b) ボストン，(c) 糖尿病．横軸を λ，縦軸を RMSE とした．弱学習器を全て用いた base の結果を青色，学習データを用いた結果を緑色，テストデータの結果を赤色で表す．

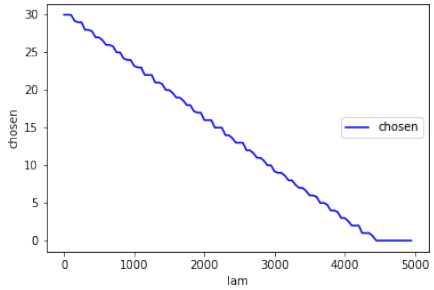


図 2: カリフォルニアのデータセットを用いて学習を行った際の選ばれた弱学習器の λ 依存性．横軸を λ，縦軸を選ばれた弱学習器の数とした．