論　　文　　概　　要

九州工業大学大学院情報工学府　　学際情報工学専攻　システム創成情報工学分野

学生番号	19677501	氏名	板橋　将之
論文題目	包除積分を利用したディープラーニングネットワークの構築		

1　序論

この研究の目的はNN（ニューラルネットワーク）を用いた包除積分モデルを構築し、学習した包除積分モデルから他の機械学習モデルとの相対的な精度の比較と、包除積分モデルから得られるシャープレイ値と XGBOOST で得られる gain や shap 値などの解釈手法と比較することを目的としている。比較対象となる機械学習モデルとしてはSVM（サポートベクターマシン）、決定木、ランダムフォレスト、XGBOOST、全結合からなるNNを使用する。

2　数学的定義

Ω＝｛1,2,…,n｝を有限集合とする。
Aを有限集合、P(Ω)をΩのべき集合、μをファジィ測度、⊗を[0,K]上のt-ノルムとする。この時Ω上の非負有界関数$f=(x_1, x_2,.. x_n) \in [0,K]^n$のμと⊗による包除積分は

$$\otimes\int f\,d\mu := \sum_{A\in P(\Omega)\setminus\{\emptyset\}} M^{\otimes}(f|A)\mu(A),$$

ただし、

$$M^{\otimes}(f|A) := \sum_{B\in P(\Omega),B\supset A} (-1)^{|B\setminus A|} \otimes_{i\in B} x_i,$$

で定義される。この式をメビウスの反転公式をを用いて変換すると包除積分は

$$\otimes\int f\,d\mu = \sum_{A\in P(\Omega)\setminus\{\emptyset\}} \left(\otimes_{i\in A} x_i\right) m^{\mu}(A),$$

ただし、

$$m^{\mu}(A) := \sum_{B\subset A} (-1)^{|B-A|}\mu(B).$$

と変形でき、この時の$m^{\mu}(A)$をエッジの重みとして学習させる。

3　提案手法モデル

包除積分をNNで構築するため図1，2のようなネットワークモデルとした。メビウス型包除積分モデル1，2では入力層から前処理層にかけて使用するネットワークが異なり、モデル1では前処理層から出力される波形は sigmoid 関数に ax+b を引数として表現される単純な波形で出力されるが、モデル2では複雑な波形も表現できるように適当なユニット数の全結合を追加したモデルとなっている。この二つのモデルを提案手法モデルとして実験を行う。

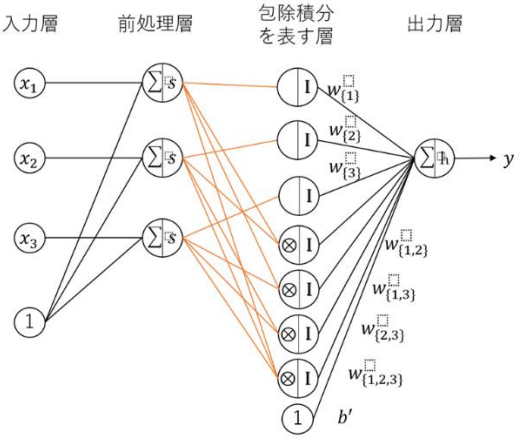

図1：メビウス型包除積分モデル1

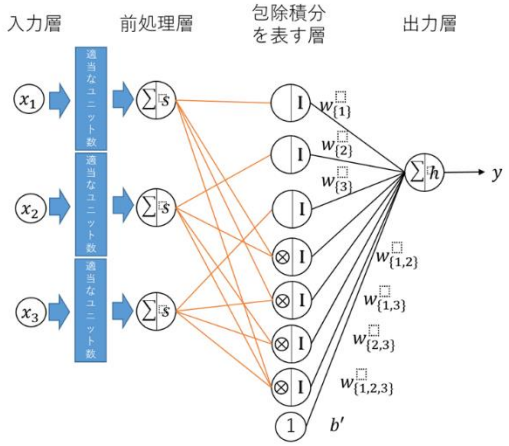

図2：メビウス型包除積分モデル2

4　実験

4.1　diabesets データの分析

実験では最初に回帰問題として、scikit-learn で提供されている「diabesets」（糖尿病）に関する 442 件のデータを使用し精度や解釈の比較を行った。目的変数は表 1 の 1 年後の疾患進行度とし、それ以外を説明変数として y の予測を行う。

実験 1 では各機械学習モデルの学習データに対する精度を測るため 442 件のデータをすべて学習データとして使用し、モデルの評価指標として平均絶対誤差、平均 2 乗誤差、決定係数の 3 点から比較を行った。

実験 2 では各機械学習モデルの汎化性能の精度を測るため 5 分割交差検証を行い、学習データとテストデータ各々に対する評価指標の比較を行った。

実験 3 では実験 1 と同様に学習を行った XGBOOST の学習モデルから gain 値と shap 値、メビウス型包除積分モデル 1，2 の学習モデルからシャープレイ値を求め比較を行い、実験 2 と同様に 5 分割交差検証で得られた XGBOOST の 5 つのモデルの gain 値と shap 値の平均を取った値と、5 分割交差検証で得られたメビウス型包除積分モデル 1、2 それぞれののシャープレイ値の平均を求めて比較を行う。また、拡張されたシャープレイ値の上位 10 を求めることさらに詳しく解釈を行った。

表 1：diabesets データの変数名

age	年齢
sex	性別
bmi	BMI 値
map	平均血圧
tc	総コレステロール
ldl	悪玉
hdl	善玉
tch	血清に関する指標
ltg	
glu	
y	1 年後の疾患進行度

4.2　Titanic データの分析

回帰問題に対する実験を行ったので残りの実験ではデータを変え分類問題に対する比較を行った。実験に使用するデータは機械学習用のデータを取り扱っている kaggle で提供されている「Titanic」データを使用した。Titanic データは文字を含むデータセットとなっているため相関が強くなるようデータの調整を施しながら数値化し、次の表 2 のようにまとめた。

表 2：Titanic データの変数

目的変数	Survived	0 or 1
説明変数1	タイトルグループ	整数
説明変数2	性別	0 or 1
説明変数3	単独乗船	0 or 1
説明変数4	旅客等級	整数
説明変数5	乗船料金	実数
説明変数6	乗船港	0 or 1
説明変数7	Cabin	整数
説明変数8	Ticket	整数

実験 4 では実験 1 と同様に精度を測り、分類問題に適用できる評価指標である正答率、適合率、再現率、f-1 値、AUC（ROC 曲線の面積の値）を追加し比較を行った。

実験 5 は実験 2 と同様に精度を測り、分類問題に適用できる評価指標である正答率、適合率、再現率、f-1 値、AUC（ROC 曲線の面積の値）を追加し比較を行った。

実験 6 は実験 3 と同様に XGBOOST とメビウス型包除積分モデル 1、2 の解釈指標を求めた。

5　結論