氏名（本籍）	下川朝有（岩手県）
学位の種類	博士（理学）
学位記番号	甲第1071号
学位授与の日付	平成27年3月20日
学位授与の要件	学位規則第4条第1項該当
学位論文題目	**On Statistical Analysis of Survival Data** **（生存データの統計解析について）**
論文審査委員	（主査）教授　宮岡　悦良 教授　石渡恵美子　教授　瀬尾　隆 教授　関川　浩　准教授　橋口　博樹

論文内容の要旨

医療データの解析において，近年の技術の進歩によりその利用できる情報は非常に増大している．一例をあげると，癌患者の持つ情報は年齢や性別といった従来から用いられてきた一般的な特徴量に加え，技術の進歩により得られるようになった腫瘍数や腫瘍の最大径とその医学的分類，また手術，放射線，化学療法の詳細な内容といった多くの情報を含む．さらにMRIやCTといった医療診断画像から得られる膨大な情報も加わる事となる．これら多くの情報の中から解析目的に合致した有意義な特徴を取り出し，正しくモデル構築を行う事は，例えば患者の処置の意思決定，また予後の予測において重要な課題となる．

本研究では，患者の予後予測の為の教師あり学習によるモデル構築に焦点を置く．特に遺伝的アルゴリズムと k-最近傍の組み合わせによるモデル構築と，Classification and Regression Tree (CART) による木構造モデルの構築について考察する．また他方の目的として，近年の技術の進歩によって得られる情報の新たな可能性の探索にも焦点を置く．より具体的に，医療診断画像から得られる特徴の可能性に着目した．

Holland (1975) によって提案された遺伝的アルゴリズムは生物の進化に端を発する最適解探索アルゴリズムの一つであり，Siedlecki and Sklansky (1989) により使用する特徴の数が 20 以上の場合，非常に有効な特徴選択の手法であることが示されている．与えられた問題に対し，このアルゴリズムはその解の候補を染色体と呼ばれる二値データで表し，解の候補の集団に対する評価と一定の操作を繰り返す事で，最適な解を探索する．従来の

特徴選択に対するこのアルゴリズムの適用では，任意の特徴をモデルに含めるか否かを染色体の一つの要素で表す．すなわち，$\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_p)$をモデル構築に考慮する p次元特徴ベクトルを表すとすると，染色体は$\{f_1, f_2, \cdots, f_p\}$で表現され，その各要素は以下で与えられる：

$$f_j = \begin{cases} 1 & Z_j\text{をモデルに含める} \\ 0 & Z_j\text{をモデルに含めない} \end{cases}$$

$j = 1,2,\cdots,p.$

この遺伝的アルゴリズムを用いて，最適な特徴の組み合わせだけでなく，各特徴量のモデル内における最適な重みの探索を行う研究が Punch et al. (1993) によって行われている．彼らは各特徴について複数の要素を用いて染色体を表現する事で，最適な重みの探索を試みた．すなわち，各染色体は$\{f_{11}, f_{12}, \cdots, f_{18}, f_{21}, f_{22}, \cdots, f_{28}, \cdots, f_{p1}, f_{p2}, \cdots, f_{p8}\}$で表現され，任意の特徴$Z_j$の重みは要素$\{f_{j1}, f_{j2}, \cdots, f_{j8}\}$を 10 進法表現で表す事で与えられた．

本研究ではこの枠組みとは異なる新たな重み探索の手法を，アンサンブル法を参考に提案する．すなわち，学習用データをランダムに二群に分け，一方を用いて遺伝的アルゴリズムによる最適な特徴の組み合わせの探索，他方を用いて評価を行う．この操作を複数回反復する事で得られる各特徴の選択される割合を重みとして扱う．

従来の重みを考慮しない手法，Punch et al. の重み付け手法，そして提案手法について，本稿では肺癌患者の悪性度の予測モデルの構築を通して，比較研究を行った．用いる特徴としては，各患者の CT 画像から得られるテクスチャー特徴に着目し，グレーレベル同時生起行列及びウェーブレット変換を用いた．画像から得られるこれらの特徴のテクスチャーに関する分類の有効性は，Dhawan et al. (1996), Laine and Fan (1993) で示されている．得られた結果から，提案手法は分類精度の向上の可能性を持つこと，そして遺伝的アルゴリズムが持つ一つの弱点である，得られる結果の不安定さ(instability)への一定の効果を持つことが示唆された．

一方，Breiman et al. (1984) によって提案された CART は，与えられた学習用データを用いて効率的に二分木構造の予測モデルを構築するアルゴリズムである．このアルゴリズムは，与えられたデータに対し再帰的にそのデータを分割する「分割」ステップと，分割ステップにより得られた木構造モデルから最適なサイズの木構造モデルを得る「枝刈り」及び「選択」ステップから成る．データが連続型の応答を持つとき，構築されるモデルは回帰木と呼ばれ，そのモデルは未知の新たなデータに対し，予測の平均二乗誤差を最小にするように構築される．すなわち，特徴ベクトル$\boldsymbol{Z}$を持つ患者の連続値を取る真の応答をX，また木構造モデルTにより与えられる予測を$T(\boldsymbol{Z})$とするとき，以下の値を最小にするように構築される：

$$R^*(T) = E\left[\left(X - T(\boldsymbol{Z})\right)^2\right]$$

一方，データが打ち切りを含む生存時間の応答を持つとき，Gordon and Olshen (1985) によって初めて CART による木構造の構築が行われた．この状況下で得られる木構造は生存木と呼ばれ，多くの研究者によって様々な木構造構築のための基準が与えられてきた．本研究ではその中で特に，シミュレーションによる比較研究により有効なことが示された 3 つの基準を用いた．1 つ目は，Davis and Anderson (1989) によって用いられた，指数モデル下での負の最大対数尤度を用いる基準であり，木構造は分割により以下の基準を最小にするように構築される：

$$R(t) = \sum_{i \in \mathcal{L}_t} \delta_i - \sum_{i \in \mathcal{L}_t} \delta_i \log\left(\frac{\sum_{i \in \mathcal{L}_t} \delta_i}{\sum_{i \in \mathcal{L}_t} x_i}\right)$$

ここで$\mathcal{L}_t = \{(x_i, \delta_i, \boldsymbol{z}_i); i = 1,2,\cdots, N_t\}$ は木構造内の分割点tに含まれる学習用データの部分集合を表し，$x_i, \delta_i, \boldsymbol{z}_i$ はそれぞれ標本iの観測時間，打ち切り指標（イベントデータなら 1，打ち切りデータなら 0），p次元共変量ベクトルを表す．

2 つ目は，Leblanc and Crowley (1992) により提案された，比例ハザードモデル下でのデビアンス残差を用いる基準であり，木構造は分割により以下の基準を最小にするように構築される：

$$R(t) = \sum_{i \in \mathcal{L}_t} 2\left[\delta_i \, log\left(\frac{\delta_i}{\hat{\Lambda}_0(x_i)\hat{\theta}_t}\right) - \left(\delta_i - \hat{\Lambda}_0(x_i)\hat{\theta}_t\right)\right]$$

ここで$\hat{\Lambda}_0$は全学習用データを用いて得られる累積ベースラインハザードのネルソン・アーレン推定量を表し，$\hat{\theta}_t$は分割点tに含まれるデータ$\mathcal{L}_t$ が従う比例ハザードモデルにおける定数パラメータの one-step 推定量を表す．

3 つ目は，Ciampi et al. (1988), Segal (1988), Leblanc and Crowley (1993)によって考察された，Log-rank 検定統計量を用いる基準であり，木構造は以下の基準を最大にするように分割され，構築される：

$$G(t) = \frac{\sum_{i \in \mathcal{L}_t}(d_{Li} - e_{Li})}{\sqrt{\sum_{i \in \mathcal{L}_t} v_{Li}}}$$

ここでd_{Li}は分割点tの分割により得られた一方の部分集合t_Lに含まれるイベント数，e_{Li}, v_{Li} はそれぞれd_{Li} の期待値と分散を表す．

本研究ではこれら 3 つの基準を用いて，乳癌から脳への転移患者の生存に関する予測モデルの構築を行った．用いる特徴としては，脳転移患者の予後予測に広く用いられている年齢や Karnofsky Performance Scale (KPS)といった共変量及び，MRI 画像から得られるテクスチャー特徴を利用した．生存時間解析において広く用いられる Cox 回帰による解析結果(Cox (1972))，及び生存木による解析結果から，脳転移患者の予後予測における画像利用の有効性，および木構造モデルの解釈の容易さにおけるメリットを示した．

Breiman らの研究，また広く提案されている木構造を用いた研究では，モデルを構築す

る際，各ノードにおける分割ルールは1つの変量によって定まる．すなわち，R^pをp-次元特徴空間，$R_t \subset R^p$を分割点tに含まれるその領域とするとき，その分割ルールは以下で与えられる：

$$"Z_j \in R_t?"$$

ここでZ_j が連続型の場合，その分割ルールは以下で与えられる：

$$"Z_j \leq s?"$$

ここでsは分割閾値を表す定数である．一方Z_jがカテゴリカル型で，その取り得る値が $F = \left\{f_{j1}, f_{j2}, \cdots, f_{jC_j}\right\}$ で与えられるとき，その分割ルールは以下で与えられる：

$$"Z_j \in R_{tj}?"$$

ここで$R_{tj} \subset F$.

しかしながら，分割ルールが 1 つの特徴によって定まるこの仮定下では，XOR 問題等に代表される幾つかの問題に対し，得られる木構造は脆弱な結果を返すことが知られている．この問題は予後をモデル化する生存木の構築においても同様に存在すると考えられる．そこで本研究では，特徴の組み合わせを考慮した分割ルールの下で生存木の構築を行う事を考察した．このとき分割ルールは，K個$(K \leq p)$ の特徴を用いて以下で与えられる：

$$"Z_{j_1} \in R_{t_1}?" \cap "Z_{j_2} \in R_{t_2}?" \cap \cdots \cap "Z_{j_K} \in R_{t_K}?"$$

$j_k = 1,2,\cdots,p,\ k = 1,2,\cdots,K$．シミュレーションによる比較研究を通して，特徴の組み合わせを考慮した際に得られる木構造のパフォーマンスは，XOR 問題下において従来の手法より遥かに良い結果を示す事を示した．また骨髄移植を受けた白血病患者のデータ解析を通して，得られるモデルは従来の手法と同様に解釈が容易であるというメリットを要する事を示した．

一般的なデータ解析において，p個の確率変数に関するデータはp-次元空間上の単一の点として与えられる．しかしながら幾つかの状況において，データを単一の点で与える事が難しい状況が存在する．例えば，ある患者の血圧のデータを取り扱いたいとき，血圧は上下する為，単一の値で与える事は難しい．また，毎日測定した患者の体温をデータとして扱いたいとき，体温は日によって異なる為，やはり一つの値として定めることは困難である．このようなデータに対し，例えば平均値のみを利用してデータの解析を行う事は，その変動を無視する事になってしまう．

そこで近年，このようなデータに対し，シンボリック変数を持つデータとして扱う手法が提案されている．これは連続値を取る変数Yに対し，ある患者iの実現値y_iを以下で与えると定義する：

$$y_i = [a_i, b_i]$$

ここで，a_i は患者i のY に関して得られた観測値の最小値，またb_i は最大値を表す．

本研究ではデータがシンボリック変数を持つときの新たな回帰木構築の手法を提案し

た．従来の手法と大きく異なる点として，提案手法では各患者のデータは複数の分割点に同時に含まれる事を許可した．この仮定下でモデルの構築を行うとき，幾つかの問題点が浮上するが，本研究ではそれらに対する対処法について提案を行い，シミュレーションによりそのふるまいの比較検証を行った．また HIV-1 型感染患者のデータを参考にした適用例を通して，提案手法の実用例を示した．

論文審査の結果の要旨

生存データの統計的解析において，患者の予後を予測するモデルの構築は重要なテーマの１つである．本論文では教師あり学習に基づくモデルの構築及び，医療診断画像を用いた予後の予測について取り上げている．

第１章は序論である．本論文の主要なテーマである，教師あり学習，生存時間データ，区間型変数を持つシンボリックデータ，医療診断画像の処理のそれぞれについて，記述及び一般的な解析法について述べている．また本論文において取り扱う実データについて述べている．

第２章では遺伝的アルゴリズムと k-最近傍法による，CT 画像を用いた肺癌患者の悪性度の分類について取り上げている．画像から得られるテクスチャー特徴に注目し，グレーレベル同時生起行列，及びウェーブレット変換を用いて画像の持つテクスチャー特徴量を数値化し，共変量として用いている．また，遺伝的アルゴリズムの不安定性に対処する為，アンサンブル法を参考に新たな手法の提案を行っている．実データの解析を通して，画像から得られる情報を分類に用いる事の有効性及び，提案手法による分類精度の向上を示している．

第３章では続く章で用いられる Classification and Regression Tree(CART)による木構造の予測モデルの構築について解説を行っている．特に応答が連続な変数で与えられる回帰木の構築及び，打ち切りを含む生存時間で与えられる生存木の構築について詳しく述べられている．

第４章では転移性脳腫瘍患者の予後予測について取り上げている．年齢や性別といった一般的に用いられている共変量と，MRI 画像から得られるテクスチャー特徴に注目し，それらを予後予測に用いる事の性能の比較を Cox 回帰及び，生存木を用いて行っている．実データの解析を通して，画像から得られる情報を予後の予測に用いる事の有効性を示している．また，木構造のモデルが与える解釈の容易さについても考察を行っている．

第５章では生存木の構築における，分割ルールについて考察を行っている．一般的な木構造のモデルでは，各層で一つの共変量のみを利用している．しかしながらこの制限下では，幾つかの問題に対しモデルが脆弱な結果を返す事が知られている．そこで本論文では，各層で複数の共変量を利用してモデル構築を行う手法を提案している．シミュレーション研究を通して，脆弱な結果を返す代表的な問題の一つである XOR 問題に対

し，提案手法の有用性を示している．また，骨髄移植を受けた白血病患者のデータ解析を通し，その解釈について述べている．

第６章では区間型変数を持つシンボリックデータに対する，新たな回帰木の構築手法を提案している．従来のデータ表現では，各データは，応答及び共変量空間上の単一の点で与えられる．しかしながらシンボリックデータの表現では，各データは空間上の超平面で与えられる．このようなデータを用いた回帰木の新たな構築手法として，各データが木構造内の複数の節点に同時に含まれる事を許容するモデルを提案し，モデル構築における幾つかの問題点への対処法を示している．シミュレーションによりそのふるまいの比較を行い，また HIV 感染患者のデータを参考に，適用例を示している．

第７章はまとめである．

以上のように，本論文は，生存木を中心として生存解析の重要な結果を与えており，医薬データ解析における統計の理論に対する大きな貢献と考えられる．よって，本論文は学位（博士）論文として十分価値があると認める．