様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　５　月　２６　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2012～2014
課題番号：　２４７００２７６

研究課題名（和文）Ｃ－ｉｎｄｅｘを分岐規準としたＳｕｒｖｉｖａｌ　Ｔｒｅｅの開発と臨床医学への応用

研究課題名（英文）Development of C-index-based survival tree and its application to biomedical research

研究代表者

林　賢一（Hayashi, Kenichi）

大阪大学・医学（系）研究科（研究院）・助教

研究者番号：７０６１７２７４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：事象が発生するまでの時間を応答とする統計モデルの評価指標であるC-indexについて，Survival Treeへ適用するための研究を行った．C-indexは，二値データにおけるROC曲線下側面積の拡張として捉えることができる量である．しかし，二値データにはない打ち切り例の存在により，統計量の単純な拡張ではバイアスが生じる．本研究では，C-indexの推定量におけるバイアスを減らすための方法を，数理的側面・計算機統計学的側面から提案し，数値実験によりそれらの有用性を示した．

研究成果の概要（英文）：This project investigated a concordance probability for time-to-event data called the C-index and its applicability to Survival Tree. The C-index can be seen as an extension of the area under the ROC curve for binary data. However, the existence of censored individuals causes a bias in estimating the C-index. We proposed some estimators for the C-index which reduce the bias, and demonstrated their superiority to the existing estimators via numerical experiments.

研究分野：　統計科学

キーワード：　医学統計学

１．研究開始当初の背景

科学的根拠に基づいた医療（Evidence based medicine; EBM）の確立は，医学研究において重要な課題のひとつである．統計解析はこの課題を達成するための手段であり，実際に多くの臨床研究で利用されている．なかでも．生存時間データに対する統計解析は重要であり，事象（疾病の発症や死亡など）に与える要因の探索・評価に用いられる．生存時間データは，一般に三つ組の変数(x,y,z)で表される．ここに，x は観察時間，y は事象の観測を示す二値変数，z は共変量ベクトルである．

本研究では，Survival Tree の理論的研究を通じて，EBM や個別治療の確立に寄与することを目標とする．Survival Tree は，データを共変量ベクトル z によって再帰的に分割し，生存時間の振舞いが異なる部分集団を構成する手法である．結果は，枝別れする「樹木（Tree）」様の構造として表現することができる． Survival Tree は，時間の情報がない二値データ(y,z)を対象とする Decision Tree の拡張として提案された．Decision Tree の長所として，以下の点が挙げられる：
(a)樹木のような構造で結果を直観的に把握でき，解釈が容易である．
(b)事象発生のリスクが高い部分集団を同定でき，個別治療の根拠を提示しうる．
(c)部分集団間で，リスク要因である共変量 z の水準が異なることが表現できる．
(d)治療方針などの意思決定に貢献しうる．
これらは，生存時間データ解析の代表的方法である比例ハザードモデルでは実現が難しい．

２．研究の目的

Survival Tree の理論的基礎付けをするために三つの問題に着目する．それは(1)新しい分岐規準の提案，(2)頑健性の確保，(3)モデル選択である．これらは相互に密接に関連した，重要な問題である．(1)は，よりすぐれた Tree の分岐規準を提案することが目的である．分岐規準とは，集団をより細かい部分集団に分割させるための規準である．Survival Tree における分岐規準として，現在までにログランク検定の検定統計量やマルチンゲール残差などが提案されている（Radespiel-Tröger et al., 2003）．これらは生存時間データの特定の面を評価する規準であるため，その面のみを強調してしまう問題がある．さらに，これらの規準の特徴や使い分けの指針などは理論的に示されていない．このような問題を解決するために， Concordance index（以下 C-index）を分岐規準として用いることを考える．C-index は，二値データにおける ROC 曲線下面積（AUC）を生存時間データに拡張したものである．これを分岐基準に用いることによって，より有用な Tree を構成することが期待される．

(2)は，(1)で提案する C-index に基づく Survival Tree に頑健性を与えることが目的である．Survival Tree は様々な長所を備える一方で，データの少しの摂動によって結果が変化してしまうという短所をもつ．すなわち科学的根拠の一要件である「結果の一貫性」に欠けることを意味する．Survival Tree を頑健な手法にするには，Bagging が有用であることはすでに指摘されている（ Hothornet al., 2004 ）． Bagging（Bootstrap aggregating）はデータをブートストラップ法によりリサンプリングした疑似標本を多数生成し，それらに対する分析結果を統合する手法である．これを C-index に基づく Survival Tree に適用することで，頑健な結果を得られると期待される．

３．研究の方法

(1)：C-index を分岐規準とする Survival Tree の構築を行う．これを達成するには，二値データにおける ROC 曲線化下面積（AUC）を最大化するブースティング(Komori, 2009)を基礎とすることが考えられる．ブースティング（Freund and Schapire, 1997）は二値データに対する判別手法であり，広義には Decision Tree もブースティングに含まれる．ブースティングは加法回帰モデルの一般化として捉えることができるため，生存時間データにも適用できることが示されている．AUC を最大化するブースティングを生存時間データについて拡張することを検討する．また，提案法を既存の手法と比較し，性質・性能の評価を行う．

(2)：C-index を分岐規準とする Survival Tree に対して Bagging を適用し，本手法に頑健性を与える．Survival Tree に対する Bagging の有効な適用法は Hothorn et al. (2004)によって示されており，これを基礎として研究を行う．

４．研究成果

(A) 生存時間データにおける C-index の推定量の改善と検証

生存時間データは，多くの場合打ち切りを伴って観測される．このため，打ち切りを考慮した C-index の推定を行う必要がある．もっとも一般的な，無情報右側打ち切りのある生存時間データに対する C-index の一致推定量は既に存在する．しかし，重みとして利用する打ち切り時間の分布やモデルのパラメータを同じ標本から推定したものを利用することに由来したバイアスが実際には問題となる．この問題に対処するために，標本を C-index の推定に利用する部分集合とパラメータの推定に利用する部分集合に分割し，推定量の改善をおこなった．この方法の一般化として，クロスバリデーションを用いた推定量を提案し，数値実験によりその有用性を例証した．その結果，打ち切り時間の分布やパラメータの推定に用いる部分標本よりも，C-index の推定に用いる部分標本が大きくな

るようにしたとき，最もバイアスが小さくなることが示された．

(B) Decision Tree に関連する統計モデル

従来の Tree で許容されている And 条件による分岐だけでなく Or 条件による分岐を許容する方法・アルゴリズムや，複数のバイオマーカの組み合わせによる分岐の指標の開発について調査をおこない，それらに関連する課題を整理し，本研究の他の統計モデルに対する応用可能を示唆した．

(C) 二重打ち切りデータに対する C-index の推定量の構成

二重打ち切りデータとよばれるデータに対する C-index の推定量を提案する．二重打ち切りデータとは，一般的に見られる右側打ち切りに加え，左側打ち切りが含まれるデータをいう．このような構造のデータに対して，usable pairs の概念を利用し，C-index の推定量の構成を考えた．Usable pairs とは，観測時間の大小関係から，事象の発生時間の大小関係が同定されるような症例の組の事である．二重打ち切りデータに対しては，通常の無情報右側打ち切りにおける usable pairs だけでなく，別の usable pairs も存在する．ノンパラメトリック型の推定量は，これらの usable pairs を用いるとともに，右側打ち切りと左側打ち切りの分布によって重み付けられた形として得られる．結果として，重みを観測値に基づいて構成可能な推定量を 2 つ提案した．これらの一方は，右側打ち切りのみが存在する場合に Uno et al. (2011)の推定量に帰着し，打ち切りが存在しない場合は両者ともに Harrell et al. (1996)の推定量に帰着することを示した．また，提案した推定量の一致性を示し，数値実験により既存の推定量と比較した，有用性を実証した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

Hayashi, K. A boosting method with asymmetric mislabeling probabilities which depend on covariates, Computational Statistics, accepted on February, 2012, 27, pp.203-218.

Takai, K., Hayashi, K. Effects of unlabeled data on classification error in normal discriminant analysis, Journal of Statistical Planning and Inference, 2014, 147, pp.66--83.

Asakura, K., Hamasaki, T., Sugimoto, T., Hayashi, K., Evans, S., Sozu, T. Sample size determination in group-sequential clinical trials with two co-primary endpoints, Statistics in Medicine, 2014, 33, pp.2897--2913.

Hayashi, K. Bias reduction in estimating a concordance for censored time-to-event response, Journal of Japanese Society of Computational Statistics, 2014, 27, pp. 1--16.

Hayashi, K., Takai, K. Finite-sample analysis of impacts of unlabelled data and their labelling mechanisms in linear discriminant analysis, Communications in Statistics -Simulation and Computation, 2014, in press.

〔学会発表〕（計 11 件）

(1) Asakura, K., Hayashi, K., T, Sugimoto., T, Sozu. and T, Hamasaki. Sample size determination in group sequential trials with two co-primary endpoints, The 26th International Biometric Conference, Kobe, Japan, 2012

(2) **林賢一.** The effects of unlabeled data on a linear discriminant function for heteroscedastic normal populations**, 統計関連学会連合大会，北海道，**2012 **年** 9 **月．**

(3) **林賢一. 打ち切りを伴う生存時間データに対する C 統計量のバイアスについて，計算機統計学会，青森，**2013 **年** 5 **月．**

(4) Hayashi, K. Evaluating discriminant performance of a semi-supervised linear discriminant analysis against a supervised one for heteroscedastic normal populations, Joint Statistical Meetings (JSM2013), Montreal, Canada, 2013.

(5) Asakura, K., Hayashi, K., T, Sugimoto., T, Sozu. and T, Hamasaki. Sample size evaluation in group sequential designs for clinical trials with two continuous endpoints as co-primary contrasts, Joint Statistical Meetings (JSM2013), Montreal, Canada, 2013.

(6) Yamamto, M., Hayashi, K. Model-based clustering for multivariate binary data with dimension reduction, Joint Statistical Meetings (JSM2013), Montreal, Canada, 2013.

(7) **山本倫生，林賢一. 多変量二値データに対する次元縮約を伴うクラスター分析法，統計関連学会連合大会，大阪，**2013 **年** 9 **月．**

(8) Yamamoto, M., Hayashi, K. Clustering of multivariate binary data via penalized

latent class analysis with dimension reduction, The 6th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2013), London, UK, 2013.

(9) Hayashi, K. On the nonparametric estimation of a concordance probability for doubly-censored time-to-event response, The 26th International Biometric Conference, Florence, Italy, 2014.

(10) Yamamoto, M., Hayashi, K. Simultaneous analysis of clustering and dimension reduction for binary variables with application to biomedical data, The 26th International Biometric Conference, Florence, Italy, 2014.

(11) **林 賢一, 清水泰隆**. Consistent estimators of a concordance probability for doubly-censored time-to-event responses**,** **統計関連学会連合大会, 東京,** 2014 **年** 9 **月.**

〔図書〕（計 0 **件）**

〔産業財産権〕
○出願状況（計 0 **件）**
○取得状況（計 0 **件）**

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
林 賢一（Hayashi Kenichi）
大阪大学・医学系研究科・助教
研究者番号：70617274