科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　３４４１６
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００２８４

研究課題名（和文）無視不可能な欠測データの一般化モーメント法にもとづく解析方法の開発

研究課題名（英文）Development of Analysis Methods based on Generalized Methods of Moments for Nonignorable Missing Data

研究代表者
高井　啓二（Takai, Keiji）

関西大学・商学部・准教授

研究者番号：２０５７２０１９

交付決定額（研究期間全体）：（直接経費）　　2,200,000円、（間接経費）　　660,000円

研究成果の概要（和文）：本研究の成果は，第一には，欠測データにもとづく最尤推定量が大標本のときに，どのような性質を持つのかを理論的に調べたことである．完全データの場合とは異なる条件下で，欠測データにもとづく最尤推定量は，一致性と最尤推定量という好ましい性質をもつことを示した．第二には，判別分析を用いる際に，ラベル分けが不完全なデータしか手に入らない場合に，どのようにデータにラベルをつけるのが良いのか，またどのようにデータを用いればよいのかを，第一の研究結果を用いて明らかにした．

研究成果の概要（英文）：As the first result of my study, I found that the maximum likelihood estimator (MLE) constructed from incomplete data has desirable properties such as consistency and asymptotic normality under the different conditions to the case in which completely observed data are available. The second result is application of the first result to discriminant analysis with partially labeled data. Since the partially labeled data can be regarded as missing data, the first result can also be applied to estimation of the parameters in such discriminant model when constructing a discriminant rule. It is found that all data available to us should be used when the observations used to construct the rule is completely randomly chosen, while there are times when not all observations with or without labels should be used for the data which are chosen depending on the value of the feature vector.

研究分野：　統計科学

科研費の分科・細目：　情報学・統計科学

キーワード：　欠測データ　最尤法　漸近理論　判別分析　部分的にラベルづけされたデータ

１．研究開始当初の背景

データを収集すると，当初の意図や計画に反して観測できない値が存在することがある．そのような観測できなかった値を欠測値と言う．例を表 1 に与える．表中の？が欠測値である．欠測値がある場合に通常の統計手法を用いるためには，平均値という簡単な統計値を求める場合さえも，様々な工夫が必要となる．これは，通常の統計手法が，欠測値のないデータ（完全データ）のために開発されてきたからである．我々が出会うデータの多くは欠測値を含むため，そのようなデータへの対処方法を考える研究は，極めて現実的な要請である．更に欠測値データ解析の理論は，欠測値と見なすことができる要素を持つ分野（潜在変数モデル，判別分析モデル，因果推論，アルゴリズムの理論など)でも利用できる点で重要である．

表 1 欠測データの例

番号	y	x
1	0	140
2	1	190
3	0	100
4	?	130
	：	
100	?	200

欠測データのための理論では，Rubin (1976)による欠測を生じさせる機構（欠測データメカニズム）の分類を用いることが標準的である．Rubin (1976)は，欠測データメカニズムを，完全にランダムな欠測（Missing Completely At Random; MCAR），ランダムな欠測（Missing At Random; MAR），ランダムでない欠測(Not Missing At Random; NMAR)に分けた. MCAR と MAR の二つは，無視可能な欠測, NMAR は無視不可能な欠測とも呼ばれる．MCAR と MAR の場合は，欠測データメカニズムをモデリングしなくても，パラメタをバイアスなく推定することが可能であるため，無視できるからである．一方，NMAR の場合に，パラメタをバイアスなく推定するためには，欠測データメカニズムのモデリングが必要であり，欠測データメカニズムを無視することはできないため，無視不可能と呼ばれるわけである．

欠測データは，測定の失敗の結果として生じたものだけを指していると考えられがちであるが，必ずしもそうではない．場合によっては，費用や目的に応じて計画的に欠測を生じさせる（させざるを得ない）ことがある．例えば，表 1 が，100 人に対して健康かどうか（これをラベルと言う．y=0 のとき健康，y=1 のとき病気）と，平均血圧（x の値）を調べたものとしよう．全ての人に対して平均血圧の値を知ることはできる．しかし，病気かどうかについては，血圧以外の要因を更に調べる必要があるので，y の値を知るためには費用もかかる上，x の値が高いからといって，無理に検査を強いることは倫理的にできない．このように分類を表す二値変数は一部のみ観測され，残りの変数はすべて観測されているというデータを部分的にラベル付けされたデータ（partially labeled data）という．このタイプのデータは，医学はもちろんのこと，社会科学においてもしばしば発生する．近年では，技術の発達によって安価にデータは取りやすくなったが，病気かどうかを判定するような倫理に関わることは依然として困難であるため，このタイプのデータはますます発生するようになっている．

部分的にラベル付けされたデータに基づいて，y=0 か y=1 かを判別するためのルールを作成することは，データマイニングにおいては重要なトピックの一つである．データマイニングの世界では，ラベルのあるデータだけでなく，ラベルのないデータも用いて，判別ルールを作ることが頻繁に行われている．このように，ラベルのあるデータとないデータの両方を使い判別ルールを構築することは，データマイニングでは半教師あり学習と呼ばれている．半教師あり学習は，ラベルのあるデータだけから判別ルールを構成するよりも，判別誤差が少ないという意味において，効果的であるという考え方が研究者の間で広く受け入れられている．そして，この考え方にもとづいて様々な方法が提唱されている．しかし，本当に半教師あり学習が効果的であるのか，ということについてはあまり研究されてこなかった．

半教師あり学習のデータは欠測データそのものであることから，欠測データ解析の知見を用いることによって，半教師あり学習が本当に効果的なのかについて研究を行うことができる．報告者は，このような考えのもと，欠測データ解析の数学的基礎についての研究（論文[2]）および，それを用いた判別分析における半教師あり学習の効果の程度についての研究（論文[1]）という，一連の研究を行った．

２．研究の目的

本研究の究極の目的は，部分的にラベル付けされたデータにもとづく判別ルールの構成が本当に判別誤差を減少させることができるのかを欠測データ解析の知見を用いて調べることである．

この目的のためには，第一に，欠測データ解析における数学的な道具を用意しておく必要がある．具体的には MCAR，MAR，NMAR によって発生した欠測データを用いたときに，最尤推定量がどのような性質を持つのかを調べることである．完全データにもとづく最尤推定量は，いくつかの数学的条件のもとで，一致性と漸近正規性という二つの重要な性質を持っていることが知られている．一致性とは，データのサイズが大きいと

き，推定量が，パラメタの真の値に十分に近いう性質であり，漸近正規性とは，データのサイズが大きいとき，推定量の分布が正規分布と見なせるという性質である．この二つの性質があることによって，検定を行うことができ，さまざまな状況で最尤推定法が用いられているわけである．本研究では，どのような数学的条件のもとで，欠測データにもとづく最尤推定量が，一致性を持つのか（あるいは持たないのか），そして漸近正規性を持つのか（あるいは持たないのか）について調べることが目的である．

第二に，この結果を用いて，部分的にラベル付けされたデータにもとづく判別ルールの構成が，本当に判別誤差を減少させることができることを調べる．その際に重要なことは，部分的にラベル付けする際のデータが，MCAR，MAR のうちのどちらのメカニズムによって選ばれているのか，という欠測データ解析の観点を取り入れることである(NMAR の場合は第一の研究から非常に扱いにくいことが分かったので本研究では取り扱わない). これはデータマイニングにおける半教師あり学習の研究では，全くと言っていいほど考えられてこなかったが，実際のデータを考える上で非常に重要な観点である．もう一つの重要な観点は，ラベルのないデータを使うのかどうかということである．ラベルのあるデータだけを使っても判別ルールを構成することができるので，ラベルのないデータを併せて使うことで，判別誤差を減少させることになるのかどうかということについて調べるということである．従って，本研究においては，欠測データメカニズムの種類（MCAR か MAR か）と，ラベルなしデータの有無（使うのか使わないのか）という合計 4 通りの場合について，判別誤差の減少について調べることが目的である．

３．研究の方法

（1）第一の研究（欠測があるときの最尤推定量の性質について；論文[2]）

欠測があるときの最尤推定量の性質の導出は次のように行った．まず，形式的に式を展開し, MCAR と MAR の下での最尤推定量の一致性の証明を行った．次に，同じように形式的に漸近正規性の証明を行った．ここで，形式的と言っているのは，数学的正当性はないが, 何らかの条件（この時点では不明な条件）のもとであれば，成立するということである. 従って，厳密な証明を行うためには，この形式的な式展開がどのような条件下で成立するのかについて探らねばならない．そして，その条件はできるだけ一般的であることが望ましい．本研究では，形式的な式展開を行い，そのための数学的条件を探索し，その条件の妥当性の検討を行った．また，その条件をより一般化した．

NMAR の場合も MCAR や MAR の場合と同じ手順で行った．式展開のための一般的な条件の導出と，その条件下での一致性と漸近正規性の証明を行った. MCAR や MAR の場合との違いは, NMAR の場合には欠測データメカニズムをモデリングする必要があるということである．そしてモデリングするということは，本来自分が興味あるパラメタではないパラメタ（局外パラメタ）の推定を行わねばならないということである．本研究では，局外パラメタが一致推定可能であるという設定の下で，最尤推定量の性質について調べた.

（2）第二の研究（半教師あり学習の効果手について;論文[1]）

本研究では，第一の研究によって得られた欠測データの最尤推定量の性質を使って式展開を行っている.

まず従来の研究において，部分的にラベル付けされたデータがどのような機構でラベル付けされているのかを文献から調査した. その結果，大半が明示してはいないが MCAR のメカニズムを仮定していることが分かった. またいくつかの研究では, MCAR の下では, ラベル有りデータだけでなく，ラベルなしデータも使う方が良いのだという前提で議論が行われていた.

本研究では，まず部分的にラベル付けされたデータを作る主要な方法として MCAR にもとづく方法と MAR にもとづく方法の二つがあることを指摘した．そして，どちらの場合にも，ラベル有りデータだけで判別ルールを構成しても，ラベル無しデータも使って判別ルールを構成しても，パラメタは一致推定できること，そして漸近正規性があることを示した．次に，ラベル有りデータだけの場合の誤判別率と，部分的にラベル付けされたデータの場合の誤判別率とをそれぞれ分母，分子にした指標を比較のために定義した．この指標は MCAR の場合には，手計算によって式の形を導くことができた. しかし, この指標は MAR の場合には陽に表すことができないことも分かった．MAR の場合の誤判別率の比較はコンピュータシミュレーション法を用いて行った.

４．研究成果

（1）第一の研究（欠測があるときの最尤推定量の性質について）

本研究によって, MCAR と MAR のもとであれば，一般的な分布の下で，最尤推定量は一致性と漸近正規性という好ましい性質を持つということがわかった．また，そのための数学的な条件についても新たに導出した.

MCAR と MAR の場合の最尤推定量が持つ性質は，完全データの場合の最尤推定量と全く同じである. 重要な違いは, その精度, つまり，分散の大きさである．欠測データの場合には, 完全データの場合よりも分散が必ず大きくなってしまう.

NMAR のときには, 欠測データを正しくモデリングすることによって，やはり最尤推定量

は一致性と漸近正規性を持つことが分かった．大きな違いは，一般に分散が，完全データの場合に比して大きくなることである．このNMARの場合の結果から,MARの場合でも(本来は不要な）欠測データメカニズムのモデリングを行うことで，推定の精度を高めること（分散を小さくすること）ができることが分かった．

（2）第二の研究（半教師あり学習の効果手について）

本研究の結果，MCARの場合にはラベル無しデータを投入しても誤判別率が改善しないことが明らかとなった．一方，MARの場合には，ラベル付きのデータの選び方によってラベル無しデータを用いた判別ルールが誤判別率に及ぼす影響が，プラスにもマイナスにもなり得ることが分かった．つまり，場合によってはラベル無しデータを用いると誤判別率を減らすことができることがある一方で，誤判別率を増やしてしまう場合もあるということである．本研究では，どのような場合に誤判別率を減らすことができるのか，あるいは増えてしまうのかについてのある程度の指針を得ることができた．より明確な指針を得るために，さらなる研究を現在行っている．

本研究の結果は，データマイニングの分野において蔓延している「ラベルなしデータでも使った方がよい」という考え方を否定する根拠になり得るものである．そして，部分的にラベル付けされたデータがどのようなプロセスによって作られているのかをきちんと見極め無い限り，ラベルなしデータを分析に投入することは手間が余計にかかるというだけでなく，手間をかけた結果として分析結果の精度を悪化させかねないという重要な示唆を得ることができた．

本研究の結果は　経験尤度を用いた場合の半教師あり学習をロジスティック回帰の場合に拡張することができている（学会発表[1]）．

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　2　件）

[1] （査読有り）Keiji Takai and Kenichi Hayashi (2014). “Effects of unlabeled data on classification error in normal discriminant analysis." Journal of Statistical Planning and Inference, 147, 66-83. Elsevier.

[2] （査読有り） Keiji Takai and Yutaka Kano. (2013). “Asymptotic inference with incomplete data." Communications in Statistics: Theory and Methods,42,3174–3190. Taylor & Francis.

〔学会発表〕（計　3　件）

[1] Keiji Takai. (2013 年 8 月 6 日). “Estimation of logistic regression parameter with partially labeled data." Joint Statistical Meetings 2013. Montreal, Canada.

[2] Keiji Takai and Kenichi Hayashi. (2012 年 8 月 1 日). “Effects on labeling mechanisms on classification error in linear discriminant analysis." 36th Annual Conference of the German Classification Society on Data Analysis, Machine Learning and Knowledge Discovery. Hildesheim, German.

[3] Keiji Takai. (2012 年 7 月 3 日). "Estimation and use of mean under monotone missingness." The 2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting (ims-APRM2012). Ibaraki, Japan.

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等 なし

６．研究組織
(1)研究代表者
高井　啓二（TAKAI, Keiji）
関西大学・商学部・准教授
研究者番号： 20572019

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者

(　　)

研究者番号：