

審査の結果の要旨

氏名：倉 橋 一 成

本研究では、疾病発症の予測やがん種のサブタイプ予測など、2値で起こる現象を予測するために多数の統計モデルを構築し、選択するときに伴う誤分類率への選択バイアスや選択の方法についての研究を行い、下記のような結果が得られた。

1. 通常、統計モデルの性能はクロスバリデーション (CV) を行って誤分類率を評価する。CVによって計算された各統計モデルの誤分類率 (Err^{CV}) のうち最も良い統計モデルのものを Err^{CVS} と表現すると、 Err^{CVS} には選択バイアスが混入することを理論数式とシミュレーションによって示した。
2. 本研究では Err^{CVS} の検定方式を導き、選択された統計モデルが有意であるかどうかを検定する手法を提案した。この検定で Err^{CVS} が有意であれば、統計モデルの性能が十分に高いということを示唆する。
3. Err^{CVS} の検定の性能をシミュレーションによって確認し、 α エラーは名義水準 (片側検定なので 2.5%) 以下に抑えられ、 β エラーはサンプル数が十分にあれば小さくなることを示した。
4. 統計モデルの性能を誤分類率の値だけで評価するのではなく、視覚的に確認するための K-sample Plot (K's Plot) を提案した。この K's Plot によって、サンプルが増えたときの統計モデルの性能の上昇を視覚的に確認できるため、統計モデルの性能を質的に評価することができる。がんサンプルのマイクロアレイデータ、特定健康診査の実データで統計モデルを構築し K's Plot を描くと、統計モデルの予測性能の推移を確認することができ統計モデルの選択に幅が広がった。

以上、本研究では多数の統計モデル候補が存在する状況で予測性能を確認し、選択するとき起こる重大な問題について指摘・問題提起し、問題解決の方法を提案した。提案した手法の性能は理論とシミュレーションから正しいことが示された。この内容はこれまでの統計モデルの選択方法に警鐘を鳴らし、今後の統計解析にさらなる発展に大きな貢献をもたらすと考えられ、学位の授与に値するものと考えられる。