【15】

氏　　名	高原淳一（たかはらじゅんいち）
博士の専攻分野の名称	博　士（薬　学）
学位記番号	第　２２８８４　号
学位授与年月日	平成21年３月24日
学位授与の要件	学位規則第４条第１項該当 薬学研究科応用医療薬科学専攻
学位論文名	**データマイニング手法による化学物質の生分解性予測に関する研究**
論文審査委員	（主査） 教　授　高木　達也 （副査） 教　授　藤岡　弘道　　教　授　宇野　公之　　教　授　堤　　康央

論文内容の要旨

経済産業省は既存化学物質の点検事業を実施しているが，このうち実験的にその生分解性が評価されている物質はごくわずかである．すべての化学物質について生分解性を評価するには膨大な時間と費用を要するため，高精度な生分解性予測システムの構築が望まれている．近年，欧米では医薬品についても環境毒性評価に関する規制が整備され，その生産量や使用量に応じて，生分解性を含む環境毒性の評価が求められ，医薬品業界にとっても，高精度な予測が求められている．

これまで，化学物質の生分解性については，定量的構造活性相関解析（QSAR）に基づいた種々の予測モデルが構築されてきた．特に，Partial Least Squares（PLS）法は大量の説明変数を同時に取り扱っても，安定した予測モデルを構築できる手法であるため，化学物質の生分解性の予測を含む様々なQSAR解析に用いられている．しかしながら，PLS法は線形手法であるため，複雑な問題の解析には適さないことも多い．そのため，本研究では，特に易分解性物質について，より精度の良い生分解性予測システムを構築するため，種々のデータマイニング手法を適用し，その有用性を検証した．

本研究では，独立行政法人新エネルギー・産業技術総合開発機構（NEDO）プロジェクト「既存化学物質安全性点検事業の加速化」において編纂されたデータベースの中から，生分解性が測定されている既存化学物質554種を選定し，解析対象とした．各化学物質の生分解性は経済開発協力機構（OECD）のテストガイドラインに準拠して測定され，28日間の生物化学的酸素要求量（BOD）が理論的酸素要求量（ThOD）の60%以上の化学物質を易分解性物質と，60%未満の化学物質を難分解性物質と定義した．ここで，化学物質の生分解性について，難分解性および易分解性をそれぞれ0および1の2値として取り扱い，目的変数$\mathbf{y}$とした．また，各化合物の構造記述子89種を計算し，説明変数行列$\mathbf{X}$として，10-fold Cross Validationにより，各解析法による化学物質の生分解性予測能を554物質の正判別率により評価した．

まず，PLS法による予測を行った．PLS法では，**X**から抽出された潜在変数行列**T**が**y**とのモデリングに用いられる．このとき**T**は**y**との相関が大きくなるように計算される．PLS法では潜在変数の数を任意に設定できるため，Jack-knife法により，潜在変数の数を最適化した．この結果，PLS法では554物質のうち439物質（79.2%）を正しく判別することができた．本研究では，この正判別率を上回るデータマイニング手法の探索を試みた．

はじめに，非線形PLS法のひとつとして提唱されているKernel PLS（KPLS）法の有用性を評価した．KPLS法では，説明変数行列**X**が高次元特徴空間上の**Φ(X)**に非線形写像され，その空間上でPLS法が行われるが，カーネル関数により**Φ(X)**の計算は不要となる（カーネルトリック）．良好な予測モデルを構築するため，Jack-knife法を行い，カーネル関数および潜在変数の数を最適化した．最適化された予測モデルによりテストデータの予測を行ったところ，454物質（81.9%）を正しく判別することができた．したがって，KPLS法は化学物質の生分解性予測に有用な手法であることが示唆された．

次に，独立成分分析（ICA）を組み込んだPLS法（ICA-PLS法）の有用性を評価した．ICA-PLS法では，PLS法の前処理としてオリジナルデータにICAが適用され，確率論的に独立な成分行列に写像されるので，PLS法による**T**の抽出はデータの分布の影響を受けない．したがって，ICA-PLS法は目的変数**y**の予測により有用な**T**を抽出し得る．さらに，ICAでは，その前過程の主成分分析（PCA）により次元を圧縮することができる．本研究では，各主成分の分散を導き，その分散が第1主成分の分散の1%に満たない成分を解析に不要なノイズとして除去し，次元を圧縮した．ICA-PLS法においても，Jack-knife法を行い，潜在変数の数を最適化した．PLS法では2〜3個の潜在変数が採用されているのに対して，ICA-PLS法では，1個の潜在変数で良好な予測モデルが構築できると判断された．これは，次元の圧縮とすべての成分の独立化により効率的な潜在変数の抽出が実現できたことによるものと考えられた．最適化された予測モデルを用いてテストデータの予測を行ったところ，453物質（81.8%）を正しく判別することができた．この予測精度はPLS法より優れ，非線形PLS法であるKPLS法と同等であった．このことから，ICA-PLS法もまたKPLS法と同様に，化学物質の生分解性予測に有用な手段となることが示唆された．

さらに，アンサンブル学習法であるRandom Forests（RF）法の有用性を評価した．RF法ではまず，与えられたトレーニングデータからB組のブートストラップサンプルデータを作成する．次いで，各ブートストラップサンプルデータを用いて樹木モデルを作成する．分類問題の樹木モデルは，ジニ係数（GI）を用いて分岐点が定められる．RF法では，決定木の各分岐はすべての説明変数の中から選ばれるのではなく，ランダムサンプリングされた変数の中で，最も良い変数が選択される．最後に，すべての結果が統合されて，新しい予測モデルが構築される．このとき，分類問題では多数決が採用される．ここで，決定木の数を800とし，各分岐でランダムに抽出される説明変数の数を9とすることで，予測の正判別率は82.9%となり，本研究で検討した手法の中で，最も良好な予測精度を示した．RF法では，予測結果と同時に，GIにより推定された各説明変数の重要度を得ることができる．この重要度が変数選択の指標となる可能性については更なる検討を要するものの，より解析的予測モデルの構築が期待できる．したがって，RF法は高精度分類法でかつ，有用な変数選択法ともなり得るツールであることが示唆された．

このように，化学物質の構造記述子から，その生分解性を精度良く予測するデータ解析手法を探索し，有用な手法を見出すことができた．本研究で検討した手法は，大量の説明変数を同時に扱うことができるため，いずれもQSAR問題のような教師データを伴う医薬学データ解析全般に適用可能であると考えられる．

論文審査の結果の要旨

化審法制定時の国会付帯決議に基づき，経済産業省は約20,000種の既存化学物質に関して点検事業を実施しているが，このうち実験的にその生分解性が評価されている物質は約1,200種に留まっている1．すべての化学物質について生分解性を評価するには膨大な時間と費用を要するため，高精度な生分解性予測システムの構築が望まれている．特に，構造活性相関による予測は有望視されている．また，このようなシステムは既存化学物質の安全性点検を加速化するだけでなく，新規化学物質の環境への影響を予測するうえでも必要性が高い．

これまで，化学物質の生分解性については，線形多項式回帰あるいはロジスティック回帰のような多項式回帰，人工ニューラルネットワーク，経験的フローチャートおよび種々のエキスパートシステムによる予測モデルの構築が試みられてきた．しかしこれらの方法の多くは、説明変数と目的変数との関係が線形であることを前提としているため，複雑な構造活性相関問題を解析する場合，十分な予測精度を示さないことがある．そのため，申請者は，より精度の良い生分解性予測システムを構築するため，KPLS法、ICA-PLS法、RF法などのデータマイニング手法を適用し，その有用性を検証した．

まず申請者は、非線形PLS法の一つである Kernel PLS 法を適用し、オリジナルなPLS法（OPLS）に比べ、全体としての予測精度が改善されているだけでなく、易分解性化学物質に対する予測精度が大きく向上していることを見いだした。

次に、ごく最近開発された手法であるICA-PLS法を適用した結果、KPLS法と同等かそれ以上の予測精度を示した。アルゴリズム的にはKPLS法の方が単純ではあるが、ICA-PLS法では、元の説明変数の寄与度を逆算することも可能であり、データ解析に有用と思われる。

更に、最近しばしば用いられる手法であるランダムフォレスト法（RF法）を適用した結果、やはり同様にPLS法より優れた結果を示した。この方法では、説明変数の重要度を求めることが容易であり、ICA-PLS法と同様、説明変数から生分解性に重要な要素を解析することも可能と考えられる。

以上のように、本論文は、化学物質の生分解性予測に対し重要な寄与を行っており、環境薬学への貢献も大きく、博士（薬学）の学位を授与するのに相応しいものと認める。