The University of Osaka
Institutional Knowledge Archive

Title	医薬学データへのデータマイニング手法および多変量解析の適用
Author(s)	日髙, 伸之介
Citation	大阪大学, 2010, 博士論文
Version Type	
URL	https://hdl.handle.net/11094/57943
rights	
Note	著者からインターネット公開の許諾が得られていないため、論文の要旨のみを公開しています。全文のご利用をご希望の場合は、大阪大学の博士論文についてをご参照ください。

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

【22】

氏　　名	日　髙　伸之介（ひ　だか　しんのすけ）
博士の専攻分野の名称	博　士（薬　学）
学　位　記　番　号	第　２３７６５　号
学位授与年月日	平成22年３月23日
学位授与の要件	学位規則第４条第１項該当 薬学研究科生命情報環境科学専攻
学　位　論　文　名	医薬学データへのデータマイニング手法および多変量解析の適用
論文審査委員	（主査） 教　授　高木　達也 （副査） 教　授　宇野　公之　　教　授　藤尾　慈　　教　授　上島　悦子

論　文　内　容　の　要　旨

【背景】コンピュータの利用が日常化し，実験機器なども高速化・自動化が進んだことによって，研究活動によって得られる「情報」の量は膨大なものになりつつある．また，実験などによって得られた資料・データもそのままでは単なる数字の羅列にすぎない．これらを「有益な情報」に活用するためには，目的に適した情報の整理・分析を行うことが必要となる．大量な情報をいかに処理するかが研究活動の重要な課題の一つとなっている．しかし，大量かつ多次元なデータが，複雑に絡み合う情報の中から，単純に見ることの出来ない隠れた相関関係や傾向を抽出することは難しい．この問題を解決する一つとして，多量のデータから有用な情報・知識を発掘することができる「データマイニング」が注目を集めている．膨大な多変量データを解析するデータマイニング手法は一般的に多変量解析と呼ばれ，様々な分野で応用されてきている．データマイニングに対する社会の期待は大きく，技術的にまだ発展を続けているにも関わらず，現状でもある程度の理論が確立された機械学習法・データマイニング技術が，すでに産業界で実際に使われ始めている．化学や製薬分野では，化合物の分子構造と生体への生理学的影響，生理活性との相関関係へのデータマイニングの適用が盛んに行われている．医療分野においては，診断や治療は従来，個々の医師の経験に基づくところが多かったが，機械学習による診断補助技術なども開発されてきている．日々蓄積されていく膨大なデータから，有益な知識を収集することでベネフィット評価・リスク管理を行うことが可能となるためデータマイニング技術が注目され応用されている．しかしながら，データマイニング技術は未だに発展途中な部分を有しており，今後克服すべき課題点も残されている．その課題の一つが，線形問題である．多変量データでは，説明変数と目的変数が明確な線形関係であるとは限らない．医薬学分野のデータには複雑な相互関係・関連性を持つものが多数存在しており，線形多変量解析手法だけでなく，非線形のデータにも対応可能な解析手法の必要性は非常に高い．そこで，本研究では，説明変数の制限を調整したデータマイニング手法の有用性を検証した．

【手法１】化学物質と生体に対する生理作用・活性との相関を明らかにする手法として，機械学習を利用した構造活性相関解析がよく知られており，様々な解析モデルが研究されている．定量的構造活性相関（Quantitative Structure Activity Relationship ; QSAR）解析による計算機シミュレーションは有機合成化学や薬理学などの分野において，実試験の代替試験法の一つとして活用されてきている．環境生態毒性の予測アプリケーションとしては，

ECOSAR (Ecological Structure Activity Relationships) や TIMES などの予測モデルが利用されている．しかし，これらのモデルは単回帰モデル・重回帰モデルであるため，解析に用いられる説明変数には厳密な変数選択が行われ良好なモデル構築が困難になっている．そこで，説明変数の数が多い場合でも良好なモデルを構築できる手法の一つであるPLSR (Partial Least Squares Regression ; PLSR) 法と変数選択法を組み合わせたMLR (Multiple Linear Regression ; MLR) 法を用いて，化合物の環境毒性予測モデルの構築を行った．解析対象には，環境省が実施した生態影響試験のミジンコ急性遊泳阻害試験結果を用いた．各化合物の分子モデリングおよび構造記述子の算出は，統合計算化学システム MOE (Molecular Operating Environment) version 2006.08 を用いて行い，R を用いて解析を行った．

【結果 1 】環境生態毒性予測を 3 次元構造記述子と logP を用いた MLR 分析と PLSR 分析で行った．比較的単純な説明変数だけでなく，立体的な構造特徴を表現可能な説明変数を用いた予測モデルを構築した．ステップワイズ法を応用した変数選択法を組み合わせることで，従来のモデルよりも予測精度を向上させた．従来の毒性予測では，構造分類などのクラス分類で予測が行われ，少数の変数しか用いられていなかったが，本研究では従来の予測モデルでは注目されていない 3 次元構造記述子から，毒性予測に有用な記述子が得られた．

【手法 2 】心電図の QT 間隔を延長して torsades de pointes (TdP) という致死性心室性不整脈を誘発するような非循環器用薬物が臨床現場で処方され，不整脈による突然死が世界各地で発生した経緯から，日米 EU 医薬品規制調和国際会議 (International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use : ICH) では承認申請する医薬品の非臨床試験の役割を明確に記載し，特に医薬品の不整脈誘発リスクを評価できる催不整脈モデルの重要性を示している．医薬品の循環器に関する毒性は，生命維持に直接関連する重篤な副作用を引き起こす可能性があることから，*in vitro* および *in vivo* の両手法を用いた安全性薬理試験による安全性評価を実施することが義務付けられている．そのため，医薬品の研究開発では，致死性の不整脈誘発リスクについて，候補化合物の持つ QT 延長ポテンシャルを早期段階で検出することが求められている．特に心筋細胞の遅延整流 K チャネル (I_{Kr}) を抑制することが，心筋の活動電位延長・心電図における QT 間隔の延長を起こすことが知られている．I_{Kr} を形成する K チャネルサブユニット分子の hERG (human-Ether-a-go-go-Related Gene) チャネルに対する薬物の作用は，薬物性 QT 延長症候群の主な原因の一つであるため，製薬企業は創薬の早期段階からこの hERG チャネルに対する作用評価を様々な方法で行われている．しかし， K チャネル遮断能を持つ薬物を患者に投与しても， QT 間隔が延長するとは限らず，QT 延長が起こっても必ず TdP が発生するわけではない．K チャネルと同時に Ca チャネルなどの他のイオンチャネル機能にも影響するようなマルチチャネルブロッカーも報告されていることから，hERG チャネルの評価だけで催不整脈性の評価を判断することは難しい．hERG チャネルに対する阻害活性予測は，様々な予測モデルが考案されているが， PLS 回帰分析や SVM のような教師有り分類手法のものが多い．しかし，催不整脈性のようなデータでは，変数とクラス間に必ずしも線形関係があるとは限らない．そこで，化合物の持つ複数のチャネル阻害能の分類・催不整脈性の評価を非線形多変量解析法の一つである Kohonen 型ニューラルネットワーク法を用いて試みた．

【結果 2 】教師無し学習法を用いて，活性情報を使わずに 2 次元， 3 次元構造記述子のみで化合物の活性に応じた分類を行った．分類に用いた情報には，創薬研究の初期段階でも得られる情報を使い，QT 延長ポテンシャルに応じた分類モデルを構築した．QT 延長・催不整脈能の予測は単純な線形関係ではないため，教師無し学習法の適応事例の一つと言える．

【考察】化合物の構造情報から，その化合物の持つ毒性を精度良く評価する予測モデルを構築した．予測精度の低かった従来の予測モデルに対して，3 次元記述子を考慮した多変量解析モデルを提案し，予測精度を改善した．本研究で検討した各手法は，多次元の説明変数を用いた解析を行うことができるため，複雑な相互関係を取りやすい医薬学データ解析全般への応用が可能であると考えられる．

論文審査の結果の要旨

コンピュータの利用が日常化し，実験機器なども高速化・自動化が進んだことによって，研究活動によって得られる「情報」の量は膨大なものになりつつある．これらを「有益な情報」に活用するためには，目的に適した情報の整理・分析を行うことが必要となる．しかし，大量かつ多次元なデータが，複雑に絡み合う情報の中から，単純に見ることの出来ない隠れた相関関係や傾向を抽出することは難しい．日々蓄積されていく膨大なデータから有益な知識を収集することで，適切なベネフィット評価・リスク管理の実施が可能となるためデータマイニング技術が注目され応用されている．しかしながら，データマイニング技術は未だに発展途上な部分を有しており，今後克服すべき課題点も残されているため，データマイニング手法の有用性の検証実例を蓄積し，検討を行う意義は深い．医薬学分野のデータには複雑な相互関係・関連性を持つものが多数存在しており，線形多変量解析手法だけでなく，非線形のデータにも対応可能な解析手法の必要性は非常に高い．

そこで申請者はまず，データマイニング手法を利用した定量的構造活性相関 (Quantitative Structure Activity Relationship ; QSAR) 解析による化合物の環境毒性予測モデルの構築を行った．説明変数が多い場合でも良好なモデル構築が期待できる手法の一つであるPLSR (Partial Least Squares Regression ; PLSR) 法と変数選択法を組み合わせたMLR (Multiple Linear Regression ; MLR) 法を用いて，立体的な構造特徴を表現可能な3次元構造記述子と logP から環境毒性予測モデルを構築し，従来のモデルよりも予測精度を向上させた．また，従来の予測モデルでは注目されていない3次元構造記述子から，毒性予測に有用な記述子が得られた．

次に，創薬の早期段階において必要となる化合物の催不整脈能の評価は複数チャネルにおける阻害活性の同時評価が必要であり，解析時の教師データの設定が困難になるため，教師無し学習によるモデル構築が解析に有効と考え，非線形多変量解析法の一つである Kohonen 型ニューラルネットワーク法を用いて，催不整脈能の評価，分類モデルの構築を試みた．活性情報を使わずに創薬研究の初期段階でも得られる情報である，2 次元，3次元構造記述子のみで化合物の活性に応じた分類と，QT 延長ポテンシャルに応じた分類モデルを構築した．QT 延長・催不整脈能の予測は単純な線形関係ではないため，教師無し学習法の適応事例の一つと言える．

以上のように，申請者が提示した医薬学分野におけるデータ解析に対するデータマイニング手法の有効性は，今回明らかにされた知見，構築された予測モデルに留まることなく，今後，様々な同種のデータ解析に関しても有用性が期待される重要な検討であり，博士（薬学）の学位授与に値するものと判断する．