様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　１　日現在

機関番号：　１４４０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５４６００３５

研究課題名（和文）新規潜在変数型回帰分析法、PCLSの開発とその医薬学データへの応用

研究課題名（英文）An approach for eliminating chance correlations and its application to pharmaceutical data.

研究代表者

高木　達也（Takagi, Tatsuya）

大阪大学・薬学研究科・教授

研究者番号：８０１４４５１７

交付決定額（研究期間全体）：（直接経費）　　3,900,000円

研究成果の概要（和文）：　回帰分析等の教師あり学習で必ず遭遇する、偶然の相関を排除するため、MDS等の分類手法により偶然の相関を見破り、排除後、回帰を行った。この私たちの手法は、人工的に偶然の相関を有するように作成された複雑な構造のデータに対し、良好な結果を与えた。
　しかし、実際に遭遇するやや単純な構造のデータ、例えば加水分解速度では、L1正則化とL2正則化をの組み合わせ手法により、偶然の相関は、十分に見破ることができることが分かった（エステル類に限るとほぼ90%の予測率）。
　結論として、複雑な構造のデータでは分類手法の組み合わせ手法が、単純な構造のデータではL1、L2正則化の組み合わせ手法が良好な結果を与える。

研究成果の概要（英文）：　We tried to develop a novel method for eliminating "Chance correlation" descriptors which appear when supervised learning is applied. As a result, we found a combinatoric method using data classification and regression methods gave better results in the case of artificial data.
　However, we also found that the appropriate combination of L1 and L2 regularization also provided better predictability in the case of real data sets which showed simpler data structures. According to Ockham's prionciple, we adopted elastic net and similar methods to eliminate chance correlation descriptors. Thus, we found the latter combinatoric method applied for predicting hydrolyzabilities of esters, amides, etc showed the best predictability (in the case of esters, the correct classification rate was 89%), when L2 regularization was carried out after L1 one.
　Therefore, it can be concluded that the former method gives better predictability for complex data, and latter one is better for complex data.

研究分野：　計量薬学、計量化学、計算化学

キーワード：　Chance Correlation　L1 Regularization　L2 Regularization　Ridge Regression　Elastic Net　Hydrolyzability　Classification　Data Mining

１．研究開始当初の背景

部分最小二乗法(PLS)は広く医薬学分野で用いられており、応用範囲も、定量的構造活性相関、薬剤疫学、DNA アレイ解析など幅広い。本来なら不要な説明変数を含んでも、潜在変数に変換する際に寄与をゼロ近くにすることにより、不要な説明変数の影響を取り除くことができるはずであるが、実際には説明変数の選択によって結果は変化する。

このため、できる限り偶然の相関による記述子を取り除くことが重要になるが、その手法は簡単ではない。しかしながら、偶然に相関する記述子は、やはり本当に相関している記述子と異なり、特別な挙動を示すのではないかと考え、様々な手法を試みた。

２．研究の目的

当初、幾つかの教師なし学習を用いて、記述子の枠内で偶然の相関を消去する手法を試みた。モデルデータセットでは、この試みは成功したが、実際のデータに適用しようとすると、様々な障害に遭遇した。例えば、重要な記述子を消去したり、説明のつかない記述子がその代わりに残存することなどが起こり、一時暗礁に乗り上げた。

しかしながら研究の過程で L1（Lasso）及び L2（Ridge）正則化をうまく組み合わせることで、部分的にせよ偶然の相関を消去できることが明らかとなったため、オックハムの原理に従い、2 年目から、よりシンプルな後者の手法を中心に検討を行うこととした。

３．研究の方法

R のパッケージである glmnet を用い、L1 と L2 パラメーターを調整することにより、L1、L2 正則化や、Elastic Net を実現した。

データセットとしては、人工データを作成したほか、加水分解性データを文献より抽出、化学記述子＋量子化学記述子での予測を試みた。

まず、手法を確認するために、答（どの記述子が排除されるべきか）の判明している人工データを作成した。人工データはサンプル数 50、説明変数数 19 個の正規分布に基づく乱数により、かつ、幾つかの変数間の相関が高くなるように作成した。説明変数の中から 6 つの変数をランダムに選び、説明変数の一部と相関が高くなるように従属変数を作成し乱数である誤差を加えた。

次に、乱数を１万個発生させ、そのなかから偶然に説明変数と相関係数の絶対値が 0.5～0.6 となる変数と 0～0.15 となる変数を選択した。その変数を使用し、１から３５番目までのサンプルでは偶然に従属変数と相関係数が高くなり、３６から 50 番目までのサンプルではその相関係数が低くなるような説明変数を 20 番目の列に加えた。結果、サンプル数 50、説明変数数 20 のデータを作成した。詳細を、表１に掲げた。

表１　人工データの作成

	y response variable		
training set	$y_i = \sum_{p=1}^{7} \beta_p x_{ip} + \beta_0 + \varepsilon_i$ $\sigma_s = \frac{1}{5}\sigma_y$ (i=1,2,…, 35)		
test set	$y_i = \sum_{p=1}^{7} \beta_p x_{ip} + \beta_0 + \varepsilon_i$ $\sigma_s = \frac{1}{5}\sigma_y$ (i=36,37,…, 50)		
	x_1-x_7	x_8-x_{30}	x_{31}-x_{35}
	応答変数の作成に使用	乱数	「偶然に相関した変数」
training set	有意な記述子	乱数	応答変数と高い相関
test set	有意な記述子	乱数	乱数

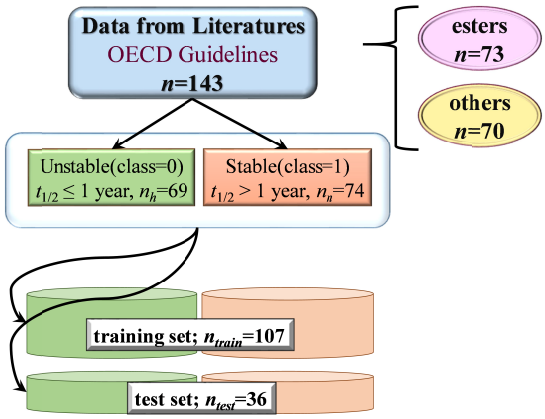

Fig.1　加水分解データセットの作成

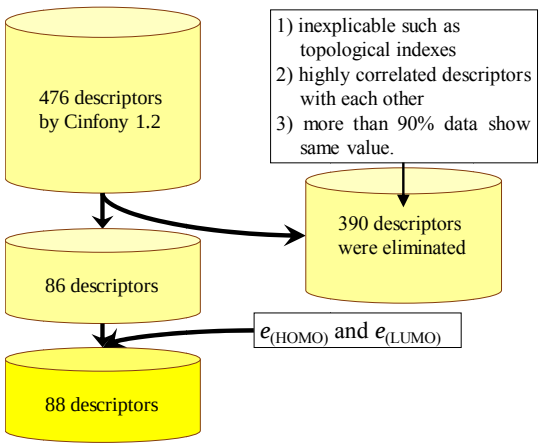

Fig.2　加水分解予測のための記述子の選択

４．研究成果

①人工データによる結果

表１のデータより、X1-X6 が実際に相関する記述子で、その他は有意でないか（X7-X19)、偶然の相関である記述子(X20)を選択した。

スケール化したデータに対して重回帰分析を行い、stepwise 法により変数を選択、予測モデル①を構築した。

次にデータをスケール化し、主成分分析を行い、varimax 回転の後、情報を持つ 20 個の主成分スコアを求めた。varimax 回転により、各主成分が説明する変数が明確となった。この主成分スコアのユークリッド距離を相関係数に基づいて作成し、クラスター分析と重回帰分析を行い、従属変数に対して寄与が小さいと判断できる主成分スコアを除去した。残った 7 個の主成分スコアを用いた部分最小二乗法（PLS）を行った。得られた結果を MLR 型のモデル式に変換し、予測モデル②を得た。

表２に各予測モデルにおける説明変数にかかる係数を示した。予測的重相関係数の自乗は、モデル①で 0.967、モデル②で 0.974 と、想定通り、モデル②の方が良好な結果を得た。

表 2　人工データによる PCLS の回帰係数

	予測モデル①	予測モデル②	実際の係数値
X1	3.199	0.206	2
X2	1.40	0.408	2
X3	-1779	-0.737	-3
X4	-2.444	-1.064	-2.5
X5	1.889	0.504	2
X6	2.016	1.205	2.
X7	-0.001	0.608	0
X8	0.049	-0.003	0
X9	-0.209	0.435	0
X10	-0.185	-0.118	0
X11	-0.187	0.389	0
X12	0.061	-0.154	0
X13	-0.135	0.211	0
X14	-0.1337	-1.231	0
X15	-1.462	0.086	0
X16	0	0.192	0
X17	1.382	-0.461	0
X18	0	0	0
X19	0	0.2	0
X20	0.0446	0.167	0

② 実データによる手法と結果

実データに対する応用例として、環境中化学物質の加水分解速度の判定予測を取り上げた。データとしては、文献[1]～[6]に記載されている化合物のうち、OECE テストガイドライン 111 に基づく試験が行われ、水中(pH = 7.0) 、温度摂氏 20～25 度において加水分解半減期が確認された化学物質(143 化合物)のみを用いた（詳細は Fig.1 を参照）

予測モデルの構築に用いた記述子は、Cinfony version 1.2 および MOPAC7 を用いて算出、Fig.2 の過程を経て、88 記述子を残した。

表 3　加水分解速度の正則化ロジスティック回帰による結果

		正解率/%（training data）			正解率/%（test data）		
		全	安定	不安定	全	安定	不安定
L1 Regularization only							
1	CV	83	82	84	69	68	72
	BIC	65	60	69	69	68	71
2	CV	100	100	100	83	66	86
	BIC	93	86	96	**89**	82	**93**
L2 after L1 Regularization							
1	CV	81	83	79	**78**	**82**	**74**
	BIC	66	75	55	64	72	62
2	CV	93	99	85	**89**	**93**	78
	BIC	88	95	77	**89**	**93**	67

表 4　正則化で残存した記述子とその内容

残った記述子	内容
ε(LUMO)	Orbital Energies of LUMO
BCUT.0	Eigenvalue based descriptor noted for its utility in chemical diversity
carbonTypes C1SP2	Characterizes the carbon connectivity in terms of hybridization
kierHallSmarts. 12	Counts the number of occurrences of the E-state fragments
SlogP_VSA3	Hydrophobic and hydrophilic effects
SMR_VSA6	Polarizability

解析方法は、当初人工データの場合と同様の手法を用いる予定であったが、検討中に、L1 と L2 の正則化をうまく組み合わせることで、最適なモデルを得られることが分かった。おそらく多くの実データでは、人工データの場合のように複雑なケースでなく、今回の加水分解速度データのように、正則化のコンビネーションをうまく利用することにより、偶然の相関を防げるのではないかと考えられる。今回は、L1 正則化でペナルティ項の重みを最適化した後、L2 正則化でリッジ係数を最

適化した場合が、最も成績が良かった。α＝0.9, 0.7 0.5 とした Elastic Net も試みたが、前者には匹敵しなかった。最終的に残った記述子のうち主なものは、表 4 に記した。残存した記述子は、加水分解に重要であることが予想されていたものばかりであり、今回の手法がうまく働いていたことを示している。

③考察

実データでは、L2 正則化と L1 正則化を組み合わせるだけで、偶然の相関を十分に取り除くことができた。事実、表 3 より、L1 正則化後に L2 正則化を行う（同時に行う Elastic Net では、思うような結果が得られなかった）ケースが、テストデータの予測を最も良好に行うことが大半であった。

先述のように、データによっては、人工データのような複雑な構造をしており、L3 正則化とでもいうべき①でとったような手法が必要になることもあると思われるが、まずは、L1、L2 正則化を駆使してみる必要があると、実データへの応用例は語ってくれているものと考える。

従来、L2 正則化は回帰係数の最適化には重要だと考えられていたが、記述子の選択にこれほど重要な働きを示すとは考えられてこなかったように思う。まだ一例だけであるため、確実なことが言える状況にはない。しかし、今後、L2 正則化が偶然の相関の除去に果たす役割について、更に考察する必要があると思われるものの、本研究計画の当初の目的はほぼ果たせたと考えている。

[参考文献]

[1]独立行政法人農林水産消費安全技術センター（FAMIC），http://www.acis.famic.go.jp/syouroku/index.htm，最終アクセス 2015/01/23
[2] 既存化学物質安全性（ハザード）評価シート，http://www.cerij.or.jp/evaluation_document/Chemical_hazard_data_02.html，最終アクセス 2015/01/23
[3] 国際化学物質安全性カード，http://www.nihs.go.jp/ICSC/，最終アクセス 2015/01/23
[4] PRTR 化学物質排出量等算出マニュアル，http://www.prtr.nite.go.jp/prtr/calc.html，最終アクセス 2015/01/23
[5]Syvain B. Lartiges and Philippe P. Garrigues, Environmental science & technology 29.5 1246-1254 (1995).
[6] OECD GUIDLINE FOR TESTING OF CHEMICALS 111 Adopted 12 May 1981

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) Tomoko Hatta, Tamaki Takaya, Tatsuya Takagi, Norihito Kawashita, Kousuke Okamoto, *Proceedings of 8th International Conference on Partial Least Sqares and Related Methods*, 2014, pp.199-200, France (**審査あり**).

2) Tatsuya Takagi, Tomoko Hatta, Norihito Kawashita and Yu-shi Tian, Tatsuya Takagi et al., *J Biom Biostat* **2016**, *7*(4,(Suppl), http://dx.doi.org/10.4172/2155-6180.C1.002 (**審査なし**).

〔学会発表〕（計３件）

1) Tomoko Hatta, Tamaki Takaya, Tatsuya Takagi, Norihito Kawashita, Kousuke Okamoto, "Development of Combinatorial Regression Method for Avoiding Chance Correlations," 8th International Conference on Parttial Least Squares and Related Methods, 26-28 May, 2014 (Paris, France).（**審査あり**）

2) Tatsuya Takagi, Norihito Kawashita, Tomoko Hatta, Tamaki Takaya, Tian Yu-shi, Kousuke Okamoto, "A Novel Combinatorial Regression Method for Avoiding Chance Correlations," XV Chemometrics in Analytical Chemistry, 22-16 June 2015, (Changsha, China) (**Key Note Lecture, Invited**).

3) 八田朋子, 川下理日人, 田雨時, 高木達也,「Chance Correlation を可能な限り回避する回帰手法の開発と応用」、第38回ケモインフォマティクス討論会、P18、10月8-9日、2015年（東京）(**審査なし**).

4) Tatsuya Takagi, Tomoko Hatta, Norihito Kawashita, Yu-shi Tian, "Predicting hydrolyzability using logistic regression analyses and regularization techniques," 5th International Conference on Biometrics & Biostatistics, Oct. 20-21, 2016 (Houston, USA) (**審査なし**).

〔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　　特になし

６．研究組織

(1)研究代表者

高木　達也（TAKAGI, Tatsuya）
大阪大学・大学院薬学研究科・教授
研究者番号：80144517

(2)研究分担者
川下　理日人（KAWASHITA, Norihito）
大阪大学・大学院薬学研究科・助教
研究者番号：00423111

(3)研究分担者
岡本　晃典（OKAMOTO, Kousuke）
北陸大学・薬学部・講師
研究者番号：70437309