4J1-3

組成情報と要素特徴量の統合に基づく化学反応量の予測

Predicting chemical reaction based on compositional information and feature descriptors

鈴木 慶介 *1　瀧川 一学 *1*3　清水 研一 *2　高草木 達 *2
Keisuke Suzuki　Ichigaku Takigawa　Shimizu Kenichi　Satoru Takakusagi

*1北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

*2北海道大学触媒科学研究所
Institute for Catalysis, Hokkaido University

*3科学技術振興機構, さきがけ
JST, PRESTO

Compositional data are a multivariate data, which includes relative information about samples, and often appears in a variety of areas, such as Geology, Biology and so on. Due to its mathematical and statistical properties, however, we can't find out any meaning through the analysis only using compositional data. This paper proposes new method to deal with this problem with appending feature descriptors of components and using appropriate transformation for compositional data to apply usual statistical or machine learning method.Finally, several experimental validations on catalysts data used in Oxidative Coupling of Methane(OCM) has also reported.

1. はじめに

自然科学において，あるサンプルの状態を構成要素の百分率や比率で表現したデータは頻繁に登場する．このようなデータ形式は組成データと呼ばれ，構成成分の相対的大小関係を表すために用いられる，その利用対象は地質学，生態学，政治学など，多岐にわたる [1][2]．しかし組成データには統計的手法を用いて解析するに当たり主に 3 つの技術的問題点が存在する．1) 要素和に定数制約が存在することから，変数間の独立性を仮定できない．例えば A:70g, B:20g: C:10g の 100g の三元混合物に対し，A を 100g，B を 25g，C を 15g に増加させると B は実際には増加しているにも関わらず，百分率表示の組成データでは B の値は減少してしまう．2) ある三元混合物 X，Y がそれぞれ，A，B，C と D，E，F で構成されていた場合，共通成分が存在しないので単純な比較は意味をなさない．3) 構成要素そのものの特徴が考慮されない．2) を例にとると，たとえ共通成分がなくとも，構成要素間の類似性を考慮出来れば統計的解析が可能である．以上より，組成データを解析する際には特殊な考慮が必要となるが，実際の応用事例では見逃されることが多い．

本研究では組成データに基づく化学反応量予測を対象とし，先行研究として触媒を構成する元素の配合比からその活性値を予測する試みが報告されている [3]．これらの触媒はメタンの酸化カップリング (Oxidative Coupling of Methane，OCM) と呼ばれる反応で用いられた触媒であり，メタンから直接他の有用資源に変換可能な反応として盛んに研究がなされている．その内容としては多元系触媒の配合比と反応量の既知データに基づいて，反応量の大きい新しい配合の触媒を予測するものであり，機械学習による予測結果と実際に計測された反応量の比較結果も議論されている．こうした試みは，発見的に行われてきた材料探索の合理化・効率化や実験コスト削減への期待から，マテリアルズインフォマティクスと呼ばれる研究領域として，近年非常に注目されている．多元系触媒の組成データに基づく活性値の予測は，触媒科学分野で先んじた事例であり注目に値するものであるが，統計手法やデータ解析の観点からは，組成データの扱いを含めて精緻化と改善が必要である．

以上の背景より，本研究では先行研究に対する技術的問題点を踏まえたうえで，同様のデータを用いてより適切かつ高精度な統計的予測手法を提案する．

連絡先: 北海道大学大学院 情報科学研究科
大規模知識処理研究室
北海道札幌市 北区北 13 条西 8 丁目
E-mail: k_ suzuki@ist.hokudai.ac.jp

2. 組成データ

組成データは構成要素の百分率や比率で表現されているデータ形式であり，自然科学において構成要素間の相対的大小関係を記述する際に登場する．しかし通常扱う実空間上のデータとは異なり，いくつかの統計的，数学的特性を持つ．本章ではその一部を述べる．

2.1 組成データの性質

組成データは割合やパーセンテージなど要素和に定数制約があるデータを指す．組成データの変数の数を D とすると，$D-1$ 個の変数のパーセント量が決定された場合，残り一つの変数の値は定数制約により一律的に決まる．つまり組成データは変数の数 D に対して，実質的な次元は 1 だけ少ないので $D-1$ 次元空間に存在する．位相幾何学的にはこの $D-1$ 次元空間を単体空間と呼び $\mathbb{S}^{D-1}$ と表記する．例えば $D=2,3$ のとき，組成データの標本空間 $\mathbb{S}$ はそれぞれ図 1(a)，(b) に示す 1-単体，2-単体となる [2]．これらは実空間と異なり有限な広がりしか持たない．

以上の性質を一般化すると以下の形で定式化される．

$$\mathbb{S}^{D-1}=\left(\boldsymbol{x}=(x_1,...,x_D)|x_i>0(i=1...D),\sum_{i=1}^{D}x_i=100\right) \quad (1)$$

組成データが有限な単体空間のみにしか存在しない性質は，通常ユークリッド空間において用いられる演算が組成データに対して不適当であることを意味する．実際，実数に対する自然な和，積は，単体空間上では閉じておらず，組成データに対する演算として不適当である点が容易に確認できる．

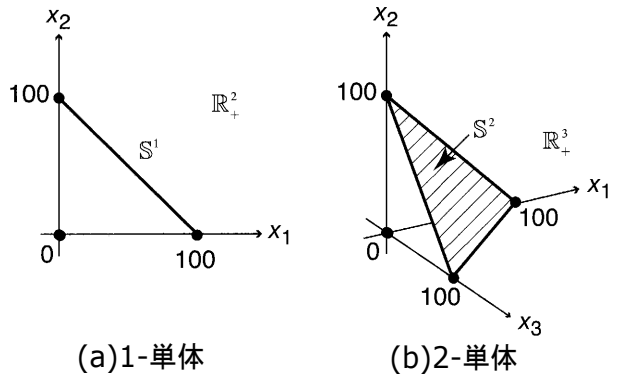

図 1: 単体空間の概念図

2.2 実空間への写像とゼロ値の扱い

組成データにはその標本空間の特性上実空間を仮定できないため，その定数制約から解放し実空間上でデータを扱うための手法が数多く提案されてきた．その中でも対数比変換は頻繁に用いられる手法である [1][4]．これは単体空間から実空間への写像であり著名なものに，相加対数比変換，有心対数比変換，等長対数比変換などが存在する．これらはすべて全単射な写像なので，実空間へ対数比変換をかけた後のデータから得られた解析結果を，逆変換で単体空間に戻すことによって，単体空間上での解析結果として確認することが可能である．この変換により組成データは定数和制約から解放されるので，実空間を仮定する様々な解析手法が適用できるようになる．

しかし対数比変換を行う場合，組成データにゼロが含まれているとそもそも対数値を定義できない．この問題を解消する方法として，ゼロ値を代替値で置換する手法が提案された [5]．

$\boldsymbol{x}$ の i 成分目 x_iをδ_i で置換することを考えると，以下の要領で $\boldsymbol{x}$ を$\hat{\boldsymbol{x}}$ に変換する．

$$\hat{\boldsymbol{x}} = \begin{cases} \delta_i, & \text{if} x_i = 0 \\ x_i(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}), & \text{if} x_i \neq 0 \end{cases} \tag{2}$$

ここで c は組成データの要素和を表す．

3. 多元系触媒の組成データを用いた反応量予測

本研究では，多元系触媒を成す構成元素の組成データから，化学反応量，すなわち触媒の活性値 (目的生成物の収率) を予測することを試みる．収率とは理論上得られる物質の最大量に対する，実際に得られた量の比であり，パーセンテージで表される．これには Kondratenko らによる先行研究が存在し，具体的には以下の形で定式化される．

ある触媒が構成要素である元素 $e_1, ..., e_m$ が，各々の組成 (配合) 比 $\boldsymbol{x} = (x_1, ..., x_m)$ で表され，かつ以下の式を満たすとき，$\boldsymbol{x}$ を D 元系触媒の組成ベクトルと定義する．

$$x_i \geq 0, (i = 1, ..., m), \sum_{k=1}^{m} x_k = 1, \psi(\boldsymbol{x}) = D \tag{3}$$

ただし ψ は，$\boldsymbol{x}$ 中の 0 でない要素の数を出力する関数である．

先行研究は，$\boldsymbol{x}_i$ を i 番目の行ベクトルに持つ行列 $\boldsymbol{X}$ を入力にとり，活性値 $\boldsymbol{y} = (y_1, ..., y_n) \in \mathbb{R}^n$ を出力として返す関数 (多項式回帰，RBF ネットワーク) を学習するものであった．$\boldsymbol{X}$ は Mg，La，Sr，Ba，Na，Cs，Li，Mn，W の 9 元素のみから成る触媒の組成ベクトルから構成されており，このとき $\boldsymbol{X} \in \mathbb{R}^{386\times 9}$ である．

しかしこの手法は，組成データの持つ統計的特性を考慮していない．実際 $\boldsymbol{x}$ は組成データの定義とは異なり 0 も含まれるので，$\boldsymbol{X}$ の列に対して疎な行列で学習を行うことになり共起成分に乏しい．また，元素そのものの特徴が $\boldsymbol{X}$ に存在しないので，元素間の類似性が損なわれている．従って本研究では以上の技術的問題点に以下の要領で対処し，統計的予測，学習精度の改良を図る．表 1 に，組成データをそのまま訓練データとして扱う際に発生する共起成分欠如の問題を具体例として載せる．例えば MgNa と MgLa の比較を行う場合，両者とも同じ Mg が構成比の大部分を占めるにも関わらず，表より La と Na が共起していないので単純な比較が意味をなさない点が見て取れる．

表 1: $\boldsymbol{X}$ の一部．異なる元系の組成ベクトルを同じ次元に統一して訓練データとしているため，共起成分に乏しい

	Mg	La	Sr	Ba	Na	Cs	Li	Mn	W
MgNa	99.0	0	0	0	1.0	0	0	0	0
MgLa	99.0	1.0	0	0	0	0	0	0	0
MgBa	99.0	0	0	1.0	0	0	0	0	0
LaSr	0	99.5	0.5	0	0	0	0	0	0
LaSr	0	99.3	0.7	0	0	0	0	0	0
MgNaMn	92.0	0	0	0	4.0	0	0	4.0	0

3.1 組成データが存在する標本空間構造の考慮

組成データが存在する空間は有限な単体空間であり，ユークリッド空間を仮定する通常の解析が適当ではなく，平均値や相関係数といった基礎統計量ですら問題が起こることが古くから指摘されてきた [6]．そのため対数比変換により実空間へ写像し，定数和制約から解放した後に訓練データとして学習することで精度の向上を図る．本研究では，2 章において触れた，単体空間から実空間への等長な写像である等長対数比変換 (Isometric Logratio Transformation，ilr)，を用いた [4]．

$\boldsymbol{e_i}$ を $\mathbb{S}^{D-1}$ の正規直交基底 $(i = 1, ..., D-1)$，$\boldsymbol{x} \in \mathbb{S}^{D-1}$ とすると ilr 変換は

$$\text{ilr} : \mathbb{S}^{D-1} \to \mathbb{R}^{D-1} \tag{4}$$

$$\text{ilr}(\boldsymbol{x}) = (\langle x_1, \boldsymbol{e}_1 \rangle_a, ..., \langle x_{D-1}, \boldsymbol{e}_{D-1} \rangle_a) \tag{5}$$

で表される．ここで $\langle \cdot, \cdot \rangle_a$ は Aitchison 内積 [1] を表し，以下の形で定式化される．

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \sum_{i=1}^{D} \ln \frac{x_i}{g(\boldsymbol{x})} \ln \frac{y_i}{g(\boldsymbol{y})} \tag{6}$$

ただし $g(\boldsymbol{x})$ は幾何平均である．また，式 (3) より ilr 変換の写像先は，単体空間 $\mathbb{S}^{D-1}$ を部分空間として内包する実空間 $\mathbb{R}^D$ ではなく，それより一つ次元数が少ない実空間 $\mathbb{R}^{D-1}$ であることがわかる．これは，D 個の変数を持つ組成データが保有する情報量が実質的に $D-1$ 個である，という特徴を反映している．

3.2 要素特徴量

組成データは，混合物同士が同じ D 元型でも共通成分がない場合，その統計的類似性を計算することに意味を見出せない．しかし要素そのものの特徴を考慮すると，同じ要素が異な

る混合物に存在してもそれぞれの比較が可能である．例えば，ある三元系触媒 X，Y がそれぞれ A，B，C，と A，B，D で構成されているとする．このとき C と D は異なる要素だがその要素の類似性を考慮できると，X，Y の部分的な類似性が考えられる．本研究では，元素そのものをいくつかの特徴で表現することで，要素間の類似性を考慮した．用いた特徴は，元素番号，原子量，族，周期，原子半径，電気陰性度，密度，イオン化エネルギーの 8 つである．

3.3 D 元型触媒に対する特徴ベクトルの構成

本研究では，扱うデータを三元系に限定し，かつ要素そのものの特徴を考慮することで，異なる要素間に共起成分が発生する様に特徴ベクトルを構成する．組成ベクトルのゼロ値を削除して，昇順にソートし直した組成ベクトルを出力する関数を ϕ，元素からそれに対応する特徴量を表す行ベクトルへの写像を desc とし，元素の要素数を．また, ベクトル $\boldsymbol{a} = (a_1, ..., a_p), \boldsymbol{b} = (b_1, ..., b_q)$ を足し合わせたベクトルを $[\boldsymbol{a}, \boldsymbol{b}] = (a_1, ..., a_p, b_1, ..., b_q)$ とする．元素の特徴量を考慮した行列 $\boldsymbol{X}_{\mathrm{desc}}$ の i 行目のベクトルは以下の形で表現できる．

$$\boldsymbol{X}^i_{\mathrm{desc}} = [\phi(\boldsymbol{x}_i), \mathrm{desc}(\acute{e}_1), ..., \mathrm{desc}(\acute{e}_{\psi(\boldsymbol{x}_i)})] \tag{7}$$

$\boldsymbol{x}_i$ は $\boldsymbol{X}$ の i 行目の組成ベクトルであり，$\acute{e}_1, ..., \acute{e}_{\psi(\boldsymbol{x})}$ は，$\phi(\boldsymbol{x})$ が表す触媒において用いられている元素である．つまり $\boldsymbol{X}_{\mathrm{desc}}$ の行ベクトルは，組成ベクトル $\boldsymbol{x}$ のゼロ値を削除し昇順にソートしたベクトル $\phi(\boldsymbol{x})$ に元素そのものの特徴を配合率の低い順から付与したベクトルになっている．

最終的に提案する特徴ベクトル表現では，両者を統合してより高精度なモデリングを試みる．この両者を考慮した行列 $\boldsymbol{X}_{\mathrm{ilrdesc}}$ の i 番目の行ベクトルは以下で表現される．

$$\boldsymbol{X}^i_{\mathrm{ilrdesc}} = [\mathrm{ilr}(\phi(\boldsymbol{x}_i)), \mathrm{desc}(\acute{e}_1), ..., \mathrm{desc}(\acute{e}_{\psi(\boldsymbol{x}_i)})] \tag{8}$$

以上の提案手法が先行研究と比較した際に有用性を持つか確認するため，実験的な検証を行う．

4. 実験

4.1 実験方法

3 説で紹介した提案手法の有用性を検証するため，以下の要領で実験を行った．

学習時における訓練データには，OCM において用いられた触媒とその活性値をまとめたデータを，テストデータには総数 42 個の触媒データを別途用意し，それぞれ使用した．表 2 に示すデータはその一部で，これらは先行研究 [3] において用いられたデータと同様のものであり，Yex は目的変数である収率 (CH_4が酸化カップリングにより C_2H_4，C_2H_6 に変換された割合) に相当する．学習には，ランダムフォレスト (RF)[7]，勾配ブースティング (GBR)[8] の二つを用いた．これらは複数の弱学習機が予測した値の多数決を取る非線形な回帰手法で，過学習を起こしにくいメリットがある．本研究ではどちらの手法とも弱学習器に決定木を使用し，それぞれのハイパーパラメータ (決定木の深さ，数) は 10Fold Cross Validation を用いた GridSearch で決定した．また GBR では，上述のハイパーパラメータに加え損失関数の種類も考慮に入れている．

学習器の精度の指標には，二乗平均平方根誤差 (Root Mean Squared Error,RMSE) を用いた．RMSE は訓練データで学習された学習器でテストデータの予測を行った際の，予測値と実測値の乖離度を表し，その値が低いほどより高精度に学習を行ったといえる．

本研究で提案した手法が持つ有用性を段階的に検証するため，以下の四条件で実験を行った．

(i):$\boldsymbol{X}(\in \mathbb{R}^{386\times 9})$ を訓練データとして学習
(ii):$\boldsymbol{X}_{\mathrm{ilr}}(\in \mathbb{R}^{386\times 8})$ を訓練データとして学習
(iii):$\boldsymbol{X}_{\mathrm{desc}}(\in \mathbb{R}^{49\times 12})$ を訓練データとして学習
(iv):$\boldsymbol{X}_{\mathrm{ilrdesc}}(\in \mathbb{R}^{49\times 11})$ を訓練データとして学習

ilr 変換により組成ベクトルの次元が一つ減少するため，$\boldsymbol{X}$ と $\boldsymbol{X}_{\mathrm{ilr}}$，$\boldsymbol{X}_{\mathrm{desc}}$ と $\boldsymbol{X}_{\mathrm{ilrdesc}}$ をそれぞれ比較すると，その列数が一少なくなる．また，本研究で提案する手法では扱うデータを三元系に限定しているため，$\boldsymbol{X}_{\mathrm{desc}}$ と $\boldsymbol{X}_{\mathrm{ilrdesc}}$ は $\boldsymbol{X}$ よりも行数（データ数) が少ない．元素の特徴量はそれぞれ主成分分析を用いて 8 から 3 に次元を削減しており，また $\boldsymbol{X}$，$\boldsymbol{X}_{\mathrm{ilr}}$ には組成情報を表す部分にゼロ値が存在するため，(2) 式に従い置換を行った．この時置換する値 δ には共通して 1^{-7} を用いた

.

表 2: テストデータの一部

	Mg	La	Sr	Ba	Na	Cs	Li	Mn	Yex
LaSrMn		90.8	9.1					0.1	14.1
LaSrBa		90.8	9.1	0.1					16.4
LaMgMn	9.1	90.8						0.1	10.3
LaMgSr	8.3	83.3	8.3						14.8
MgNaBa	90.1			0.9	9.0				1.8
MgBaCs	83.3			8.3		8.3			15.3

4.2 結果と考察

表 3 に，先行研究 [3] の結果と提案手法で得られた結果を示す．表中の数値は RMSE であり，括弧中の値は誤差の標準偏差に相当する．また先行研究の結果は，RBF ネットワーク (RBFN) を用いて学習した値になる．

また図 2，3 に，先行研究と実験環境 (i),(iv) で得られた学習器との予測精度の差を可視化した散布図を示す．この図は学習器が予測した値を水平軸，真値を垂直軸に取り散布図としてプロットしたもので，線上に多く点が集まっている学習器ほど予測値と真値のずれが少ない．

表 3: 異なる条件下における実験結果

実験環境 (訓練データ)	RF	GBR	先行研究 [3]
考慮なし ($\boldsymbol{X}$)	6.9805 (3.1556)	9.4823 (5.5346)	5.0208 (3.5671)
ilr 変換 ($\boldsymbol{X}_{\mathrm{ilr}}$)	4.7921 (3.1995)	4.4132 (2.8023)	
特徴量 ($\boldsymbol{X}_{\mathrm{desc}}$)	3.6000 (2.5452)	3.6328 (2.3407)	
ilr 変換+特徴量 ($\boldsymbol{X}_{\mathrm{ilrdesc}}$)	2.9815 (2.0585)	3.3109 (2.0851)	

表 3 より，組成情報をそのまま学習させた実験 (i) は先行研究の値を下回っているが，ilr 変換により組成データの持つ標本空間の構造を考慮した実験 (ii)，要素の特徴量を加えた実験 (iii) のどちらの場合でも，両手法とも先行研究と比べ高い精度を示した．また，組成情報と要素特徴量に対する考慮を統合した実験 (iv) では，両手法とも他条件に比べ最も高い精度を示

した．加えて図 2，3 から，提案手法を用いた実験 (iv) が，先行研究と無考慮の実験 (i) と比べより線上に多く点が分布している傾向が見て取れる．

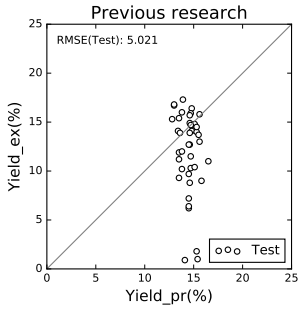

図 2: 先行研究 [3] の予測値に対する真値の分布

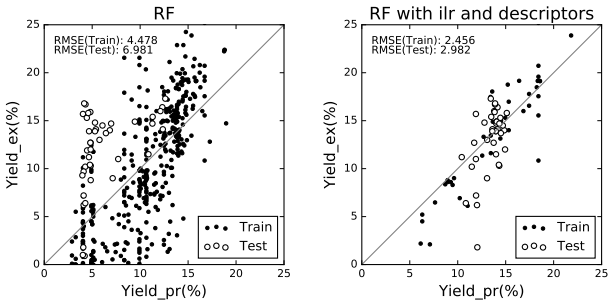

図 3: 実験環境 (i)（左）と実験環境 (iv)(右) における，予測値に対する真値の分布

5. まとめと今後の課題

本研究では，元素の比率で表現された触媒データを対象に，組成データの統計的特性と元素の特徴量を統合した学習方法を提案した．また，先行研究との比較を行い，その有用性を確認した．しかし現段階では，同じ構成要素数を持つデータ同士の比較しか出来ていない．つまり触媒を例にとると，異なる元系の触媒を同時に訓練データとして学習したとしても，データ数が膨大でない限りは訓練データの列が疎になり共起する成分が少なくなるので精度に期待はできない．従って今後の課題として，組成データの構成要素数に依らず，より汎用的に適用できる手法を考案する必要がある．また，本研究で扱ったデータの目的変数はパーセンテージで表現されているので，学習器の出力の値域を考慮した手法が求められる．

謝辞

本研究は JSPS 科研費 26330242，16K13852 および JST さきがけの助成を受けたものです．

参考文献

[1] K. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., 1986.

[2] 太田亨, 新井宏嘉. 組成データ解析の問題点とその解決方法. 地質學雜誌, Vol. 112, No. 3, pp. 173–187, 2006.

[3] E. V. Kondratenko, M Schluter, M Baerns, D Linke, and M Holena. Developing catalytic materials for the oxidative coupling of methane through statistical analysis of literature data. *Catal. Sci. Technol.*, Vol. 5, pp. 1668–1677, 2015.

[4] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, Vol. 35, No. 3, pp. 279–300, 2003.

[5] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, Vol. 35, No. 3, pp. 253–278, 2003.

[6] K. Pearson. Mathematical contributions to the theory of evolution:on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc. Roy, Soc.*, Vol. 60, No. 1, pp. 489–498, 1897.

[7] L. Breiman. Random forests. *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[8] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, Vol. 29, pp. 1189–1232, 2000.