修 士 論 文

データマイニングを用いた量子計算データからの二元合金の物性予測

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

鈴木　大輔

2014 年 3 月

修 士 論 文

データマイニングを用いた量子計算データからの二元合金の物性予測

指導教員 Dam Hieu Chi 准教授

北陸先端科学技術大学院大学
知識科学研究科知識科学専攻

1250025 鈴木　大輔

審査委員: Dam Hieu Chi 准教授 (主査)
Ho Tu Bao 教授
藤波 努 教授
Huynh Nam Van 准教授

提出年月: 2014 年 2 月

目 次

図 目 次

表 目 次

第1章　研究背景と目的

1.1　研究背景

1.1.1　国際競争のカギを握る材料開発力

我々は身の回りの製品や設備に対し、高性能化や小型化のような製品・設備自体に対するものから環境負荷軽減や省エネ化などの社会文脈に依存したものまで様々な要求を抱いており、それらの要求を満たすために先端材料・物質が使用されている。例えばレアアースは、図 1.1 に示すように次世代自動車の小型モーター、排気ガスの浄化触媒、パソコン・携帯電話の液晶研磨剤など、それぞれの製品に対する要求を満たすために使用されている。このようにレアアースは現在の産業にとって非常に重要な金属資源となっている一方、貴重かつリサイクルが困難な資源であるため安定供給が難しく、代替材料の開発や資源の有効活用が重要な位置付けにある。従って、材料開発技術を持つことは持続的社会の実現だけでなく国際競争の上でも優位であると考えられている。

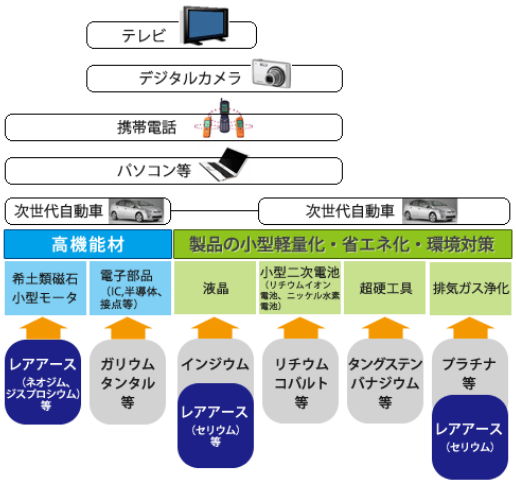


図 1.1: 経済産業省「レアアースの主な用途」

現在の材料開発は、理論・実験・計算の 3 つの軸がある。理論は現象を的確に説明するモデルの作成、実験は得られた結果の解析よる内部構造の推測、計算は実際に行うのが困難な状況のシミュレーションが行われており、これら 3 つの手法の協力関係のもと材料開発の成果が上げられてきた。

一方、これらの手法による材料開発は研究段階から実用化までに 10 年から 30 年という膨大な時間を要しており、この研究開発時間を短縮する新たな手法を開発することが大きな課題となっている。アメリカでは 2011 年から既に Materials Genome Initiative for Global Competitiveness[1] が提言されており、政府が取り組むべき重要項目の 1 つに先端材料の開発・導入にかかる時間の短縮を挙げている。日本も後に続く形で研究開発戦略センター（CRDS）が戦略プロポーザルとして立案 [2][3] し、取り組みが始まったところである。これらのプロジェクトは情報科学の視点からのアプローチが取り入れられており、材料開発において新しい試みが始まりつつあると言える。

図 1.2: Materials Genome Initiative

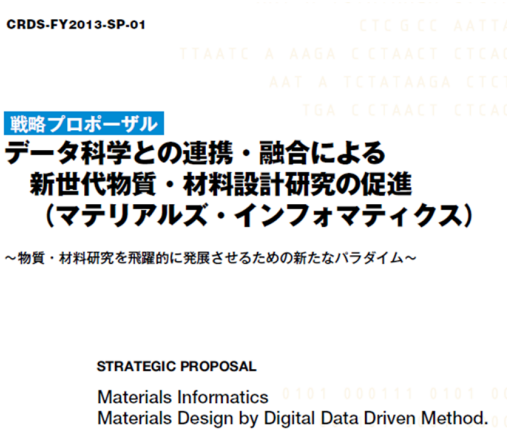


図 1.3: CRDS 戦略プロポーザル

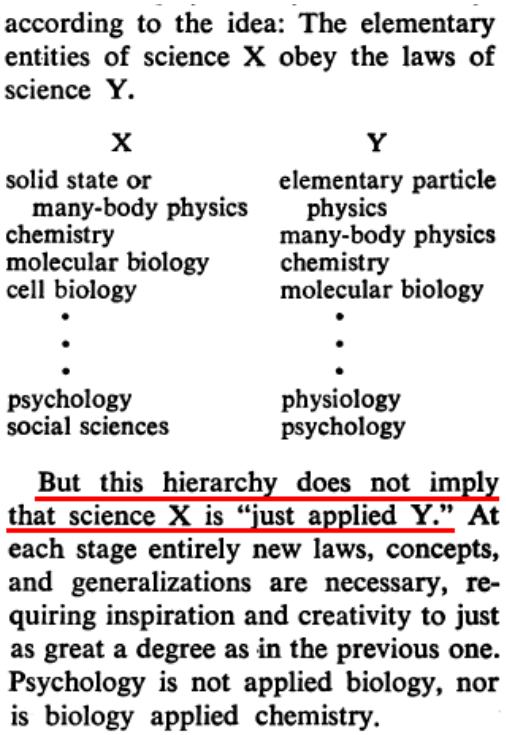
1.1.2　計算材料科学の材料設計アプローチ

計算機の高速化と低コスト化に伴い、計算科学に基づいた材料科学分野の研究開発は目覚ましい発展を遂げている。量子力学計算によるナノ材料設計、分子動力学シミュレーションによるタンパク質、高分子ポリマーの解析、半経験的計算によるデバイスシミュレーションなど、これらの技術は分析者が経験や勘を頼りに元素組成や構造などのパラメータを設定し計算することによって物性を得る演繹的アプローチであり、材料科学分野では物性予測にとって不可欠な手法である。一方、近年機械学習やデータマイニングのような、膨大なデータから帰納的に有用な情報や知識を獲得する手法に注目が集まっている。これらの手法はデータ駆動（帰納）型の知識発掘法であり、経験や勘、専門知識からでは発見が困難な知識が得られることが期待されている。

材料設計とは、どのような材料を組み合わせ、どのようなプロセスでもって目標の機能を有する材料を作成するかを計画することであり、低コストかつ短期間で生産することが必要である。材料科学分野では、データマイニングや機械学習の導入がまだそれほど行われておらず、演繹的アプローチによる材料設計が主となっている。演繹的アプローチは、条件・仮説設定、実験・計算、結果の評価の一連のプロセスを何度も繰り返し行い目標機能へ到達する方法論であるため、研究段階から応用までに多くの時間を要するという問題があるが、データから知識を自律的に採掘する帰納的アプローチと融合した有効な応用フレームワークを用いることで、特定の機能を有する材料の合理的探索の実現が期待される。その理由を、演繹と帰納による材料設計の概念と絡めて次項にて述べる。

1.1.3　演繹と帰納による材料設計

一般に多数の原子が集まったマクロな世界の原理は、少数の原子で構成されるミクロな世界の原理によって説明することが出来ない。このような現象はカオスあるいは複雑系と呼ばれており、ノーベル物理学賞を受賞した P.W.Anderson は”More is different”と表現しており、系の構成要素の数（複雑さの階層）が変わるとその系を支配する基本原理も変わってくると説明している。（図 1.4）



according to the idea: The elementary entities of science X obey the laws of science Y.

X	Y
solid state or many-body physics	elementary particle physics
chemistry	many-body physics
molecular biology	chemistry
cell biology	molecular biology
•	•
•	•
•	•
psychology	physiology
social sciences	psychology

But this hierarchy does not imply that science X is "just applied Y." At each stage entirely new laws, concepts, and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one. Psychology is not applied biology, nor is biology applied chemistry.

図 1.4: P. W. Anderson, More Is Different (1972)

材料設計において理想とされるのは、目標機能（アウトプット）を出発点として、それを実現するために必要なインプットを導き出す逆問題的設計法であると言われている。”More is different”の観点で考えると、目標とする材料物性はマクロの世界の原理に従うためミクロの世界の原理を用いて設計を行うことは困難だが、マクロの世界の原理を用いて設計を行うことは容易なはずである。従って、マクロの世界の原理を獲得すること

が、効率的材料設計の近道となる。

このマクロの世界の原理を得るために、本研究では演繹と帰納の融合手法を提案する（図 1.5）。まず、解析対象となる各物質を価電子数や電気陰性度などの実験/基礎物性と、結合エネルギーや電荷などの量子計算により演繹的に得られた物性を変数に持つデータとして表現する。そしてこれらの物質群のデータを使用したデータマイニングにより目標機能を設計する上で必要な物質群の物理を獲得する。

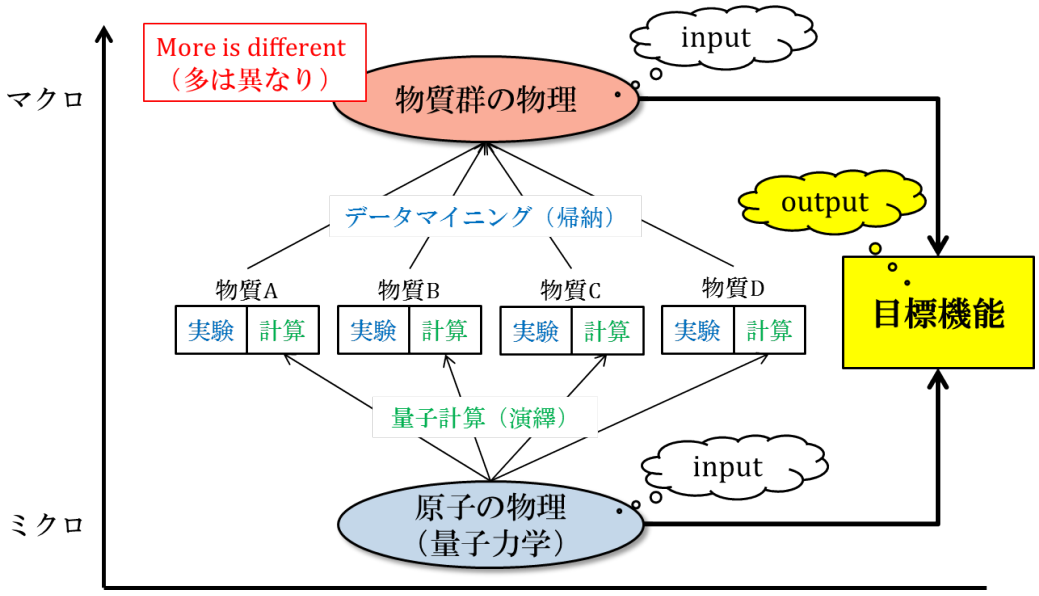


図 1.5: 演繹と帰納による材料設計の概念図

1.2　研究目的

帰納（データマイニング）と演繹（量子計算）を融合させた物性予測手法の提案が本研究の目的である。具体的な事例として二元合金の融点予測と予測モデルのグラフ化を行い、手法の有効性を評価する。尚、研究事例に二元合金を選定した理由は、合金は工業的に最も基本的な物質であり、その中でも 2 種類の原子で構成されるシンプルな構造となっているためである。同様に、基本的な物性であり、かつ合金化の目的[1]に適合した物性であることから融点に着目した。また、融点は演繹的に導出するのが難しく実験データが大量にある物性であるため、本手法の妥当性を測るに格好の対象であると言える。

1.3　知識科学的意義

本研究では計算機を利用した演繹と帰納の融合アプローチを行うが、これは人と計算機、それぞれの特性を活かした効率的融合を実現した手法であり、その有用性を評価し提案することに知識科学的意義がある。

計算材料科学において、今までの計算機の用途といえば、実験が困難な事象のシミュレーション（演繹法）が主であった。しかし、近年は計算機の高性能化・低コスト化と機械学習やデータマイニングのような解析手法の発達により、計算機の網羅性・正確性を活用したデータからの知識獲得を行うこと（帰納法）が可能となった。本研究は量子計算データに対してデータマイニングを行うが、これは研究活動における実験から評価のプロセスを計算機上で再現することに同意であり、計算機の力をフルに活用した知識獲得方法であると筆者は考えている。また、これにより人の役割は知識獲得プロセスの運転および得られた知識を元にまた新たな仮説を立てることに専念され、人と計算機の効率的融合がなされることになる。

[1]合金化により融点が低下する

1.4　本論文の構成

本論文の構成を以下に示す。

第 1 章 研究背景と目的を述べた。

第 2 章 本研究で使用するデータと主要な手法について述べる。

第 3 章 本手法の効果測定を行い結果について述べる。

第 4 章 本論文のまとめと今後の課題を述べる。

第2章　研究手法

本論文で提案する知識獲得手法は以下の流れで行われる。

1. 二元合金データベースの作成
2. 重回帰分析手法 LASSO による融点予測モデル構築
3. 融点予測モデルの拡張とグラフ化
4. グラフによる結果の理解

まず、実験/基礎物性データの収集と量子計算による物性データ作成を行い、それらのデータを統合し二元合金データベースを作成する。次に、本手法のキーアイディアである重回帰分析手法LASSOを作成したデータベースに適用することにより、融点予測モデルを構築する。さらに、融点予測モデルにおいて予測対象の物性を順次入れ替え並列にLASSOを実行し、結果をまとめ上げることにより二元合金の物性関係の全体像を獲得する。最後にその関係性をグラフで以って表現し、結果の理解を行う。本章では1. から3. について詳細を述べる。

2.1　二元合金データベースの作成

本手法においてデータ作成は非常に重要である。計算科学や統計分野で”Garbage in, garbage out”という言葉が表す通り、良いデータを使用して解析しなければ有用な知識を得ることは出来ない。本手法では解析対象データとして、実験/基礎物性データと量子計算データを統合し使用する。ここで、実験/基礎物性データは実測値であり、量子計算データは結合距離や結合エネルギーなどの実験では得難い、第一原理などの基本原理から導出された理論値である。

本研究に使用する二元合金データの作成は以下の手順で行われる。

1. 二元合金及び合金構成原子に関する実験データ/基礎物性の収集

2. 二元合金に関する量子計算データの作成

3. 実験データ/基礎物性と量子計算データの統合

まず、実験/基礎物性データを収集する。実験/基礎物性データとは二元合金の融点や構成原子の価電子数、電気陰性度、第一イオン化エネルギーなどの基礎物性である。二元合金の融点は合金相図ハンドブック Binary alloy phase diagrams から、構成原子の物性データは独立行政法人 物質・材料研究機構（NIMS）のホームページ [1] や web 百科事典 wikipedia[2] の金属元素の項目から収集した。尚、収集した実験/基礎物性データは Yousef ら [5] が二元合金の融点予測を行った際に使用した物性を参考にしている。

次に、量子計算データを作成する。量子計算データとは合金の結合距離、結合エネルギー、電荷、クーロン相互作用などの物性であり、分子モデルを構築し量子計算を行った結果得られる物性値である。尚、本研究で使用する量子計算データは全て総合的モデリング/シミュレーションソフト materials studio に収録されている密度汎関数理論に基づいた量子力学計算プログラム DMol3 を用いて作成した。量子計算データの作成手順を以下に示す。

1. 二元合金の分子モデル構築

2. 量子計算を行い分子モデルの構造最適化

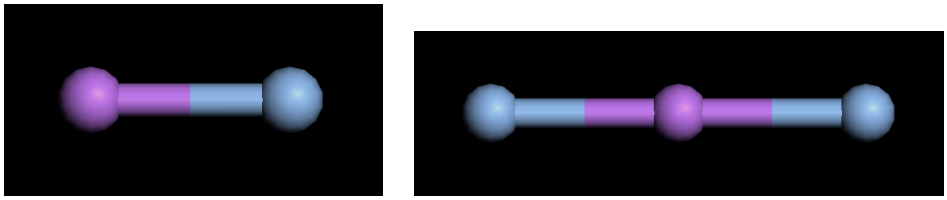
3. 構造最適化後の分子モデルを使用して物性値を計算

4. 解析に使用する物性を収集する

まず二元合金の分子モデル構築だが、本研究では、図 2.1 のような 2 つ及び 3 つの金属原子からなる分子モデルを構築した。一般的に金属結晶の最小構造単位として図 2.2 に示した面心立方格子構造 (fcc)、体心立方格子構造 (bcc)、六方最密充填構造 (hcp) が用いられるが、本研究ではより単純な構造を用いて議論を進めていく。分子モデルに使用した金属原子は原子番号 3 番以降の金属原子から選択しており、対象の元素を図 2.3 の赤枠内にに示す。また、3 つの金属原子からなる分子モデルは、原子の配置が A-B-A と B-A-B の 2 つのパターンについて準備した。

[1] http://www.nims.go.jp/

[2] http://en.wikipedia.org/wiki/Wikipedia, 収録されている情報量の関係から英語版を使用



(a) 2 原子　　(b) 3 原子

図 2.1: 分子モデルのイメージ

図 2.2: 主要な結晶構造

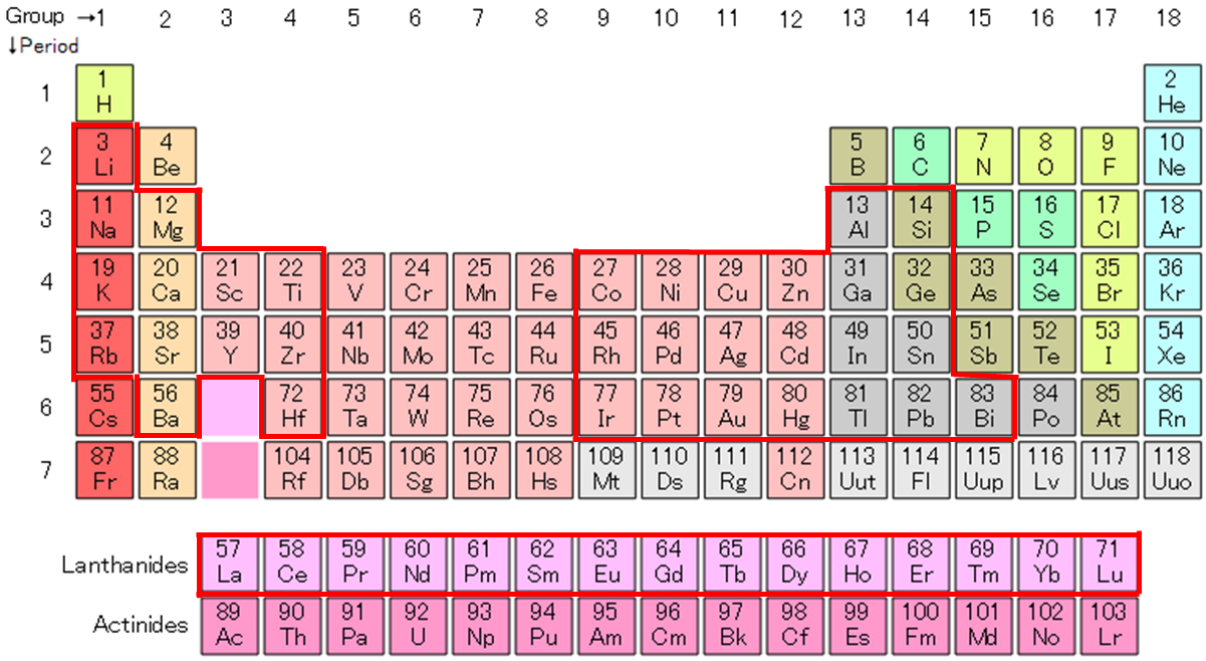


図 2.3: 周期表

次に分子モデルの構造最適化を行った。構造最適化とは、分子モデルのエネルギーが最小となる最も安定した構造を求めることである。最適化構造に基づき得られる量子計算結果を解析することで物性値に対する統一的な結論が得られるため、構造最適化は正しい量子化学計算を実行するための重要な第一段階である。本研究では構造最適化計算において、密度汎関数法に GGA（一般化勾配近似）を使用した。密度汎関数理論とは、多電子問題を 1 電子問題に書き換える基礎を与えるもので、特に LDA（局所密度近似）は個体や分子のバンド理論、凝集機構の説明に用いられ成功を収めてきた。今回使用する GGA は、交換エネルギーに対する密度勾配補正を行うことで精度を更に向上したもので、電子相関を考慮する高精度の分子軌道計算と比肩しうる計算精度を実現する一方、計算コストが圧倒的に少ない。また、GGA の関数の中でも、強い物理的背景を持ち信頼性の高い数値が得られることで知られている PBE を交換相関汎関数に選択した。

モデルの内殻電子処理は、擬ポテンシャルは使用せず、全電子の相対論的効果を計算している。これは本研究では希土類金属も合金モデルの構成要素として使用しているが、希

土類金属のような原子番号が大きくｄ電子やｆ電子が化学結合で重要な役割を果たす元素の化合物における電子状態を調べるには相対論的効果が重要になってくるためである [3]。基底関数は数値基底である DNP（double numerical with polarization）を使用した。セルフコンシステント計算の収束条件は高精度計算を目的として 1×10^{-8}Hartree[4] とした。

尚、3 つの金属原子からなる分子モデルについては、構造最適化の際に結合角 [5] が生じる事がある。これは確かに構造最適化の結果なのだが、このような 3 つの原子からなる分子が自然界に存在する場合の話である。しかし、自然界で実際にそのような状態で分子が安定して存在することはありえない。本来は無数の原子が秩序を持って並び結合しあっているはずであり、このモデルはあくまで合金構造の一部分として抜き出したものである。従って、本研究では構造最適化時に結合角が生じぬよう構造最適化の際に変位可能な範囲を x 軸方向に束縛し、3 つの金属原子が一直線上に並んだ最適化構造が得られるように設定している。

さらに、構造最適化後の分子モデルを使用して、計算条件は構造最適化時と同一のままエネルギー計算を行った。エネルギー計算の際は、モデルの調和振動数および電子密度解析によりマリケン電荷とヒルシュフェルト電荷を求めるよう設定し、得られた結果からそれらの物性だけでなく、モデルの結合距離、結合エネルギー、HOMOLUMO エネルギーギャップなどを収集した。

以上により得られた実験/基礎物性データと量子計算データを統合することにより、各二元合金を計 30 個の属性を持つデータとして表現した。[6]

$$\text{二元合金データ}：X_i = \{{x_i}^1, {x_i}^2, {x_i}^3, \cdots, {x_i}^{30}\}$$

本研究では、このような二元合金データ計 103 種 [7] を解析用データとして使用する。

[3] http://jolissrch-inter.tokai-sc.jaea.go.jp/pdfdata/JAERI-Review-99-008.pdf
[4] 1 Hartree $= 4.3597482 \times 10^{-18} J$
[5] 分子構造の構造要素の一つで、それぞれの原子から伸びている 2 つの化学結合のなす角度のこと
[6] 詳細は付録 A 参照
[7] 詳細は付録 B 参照

2.2　物性予測モデル構築

2.2.1　LASSO

前項で作成したデータに対して、目的変数を予測対象物性、説明変数を他の全ての物性に設定し線形回帰分析を行うことにより物性予測モデルを構築する。線形回帰分析は目的変数（予測対象）と説明変数（パラメータ）の間に式を当てはめ、目的変数が説明変数によってどのくらい表すことができるかを定量的に分析する手法であり、予測モデルとして線形モデルを得られるという特徴がある。線形モデルは、非線形で複雑な構造であっても説明変数の線形結合形で近似的に表現できるという利点があり、本研究では合金物性の持つ内部構造をシンプルな形で獲得するために、線形回帰分析による物性予測を行う。

最も基本的な線形回帰分析として最小二乗法があるが、最小二乗法はモデルのデータへの過剰適合を引き起こしやすいことが知られている。本研究では物性予測において重要な説明変数のみが組み込まれたモデルの獲得を目指しているため、最小二乗法はこの目的に適さない。そこで、最小二乗法に変数制御をするペナルティ項を加えたLASSO[6]と呼ばれる正則化手法と交差検定法を用いることで、必要な変数のみ選択され、かつ、過剰適合を抑えたモデルを獲得する。

最小二乗法の解は次の式を最小化することで得られる。

$$\frac{1}{m}\sum_{i=1}^{m}\left({y_i}^{predict}-{y_i}^{obs}\right)^2$$

ここで、mはデータ数、${y_i}^{predict}$は得られたモデルによる予測値、${y_i}^{obs}$は実測値を表す。また、${y_i}^{predict}$は次式のように説明変数の線形結合によって表される。

$${y_i}^{predict}=\sum_{j=1}^{n}\beta^j{x_i}^j+\beta^0$$

ここでnは全説明変数の数、${x_i}^j$はj番目の説明変数の値、β^jは${x_i}^j$に対応する回帰係数、β^0は切片である。

LASSOは最小二乗法に変数制御をするペナルティ項を加えた重回帰分析手法であり、次の式で表される。

$$\frac{1}{m}\sum_{i=1}^{m}\left({y_i}^{predict}-{y_i}^{obs}\right)^2+\gamma\sum_{j=1}^{n}|\beta^j|$$

γ はチューニングパラメータと呼ばれ、この値が大きいほどペナルティ項の効果が大きくなりモデルの回帰係数の値が縮小される。LASSO は「変数の中で真に重要なものは少数である」というスパース性の仮定のもと解を求める手法であり、回帰係数の一部がゼロとなるスパースなモデルの獲得が期待出来る。また、変数選択と係数の決定が同時に行われるため、変数増加法や変数減少法などの変数選択法と比較して計算が効率的であることが知られている。また、LASSO による予測モデルは説明変数の線形結合形で得られるため、本研究においては二元合金の持つ複雑な物性関係を、線形結合というシンプルな形で近似できるという利点もある。

ここで LASSO はペナルティ項の効果により、いくつかの係数パラメータを真にゼロと縮小することができると述べたが、その理由を 2 変数（$j=2$）の例を用いて説明する。ペナルティ項付重回帰分析の一般式としてブリッジ回帰があり、次の式で表される。

$$\frac{1}{m}\sum_{i=1}^{m}\left({y_i}^{predict}-{y_i}^{obs}\right)^2+\gamma\sum_{j=1}^{n}|\beta^j|^q$$

ブリッジ回帰において $q=1$ の場合が LASSO に該当する。$q=2$ の場合を Ridge 重回帰分析と言い、ここでは LASSO の結果との比較対照に使用する。

LASSO と Ridge の $j=2$ の場合における、誤差項の等高線表示 (楕円) とペナルティ項領域を図 2.4 に示す。Ridge ではペナルティ領域が円形 (左)、LASSO では四角形 (右) となっているのが分かる。この誤差項領域とペナルティ項領域の接点が解として得られるモデルの回帰係数 $\{\beta_1,\beta_2\}$ となるが、Ridge の場合は接点が円周上になりやすく $\{\beta_1,\beta_2\}$ 共にゼロ以外の値を持つことが多い。一方 LASSO の場合は接点がペナルティ領域の角（軸上）になりやすいため、回帰係数を真にゼロへと縮小することが出来るのである（図 2.5）。

[8]https://onlinecourses.science.psu.edu/stat857/node/158

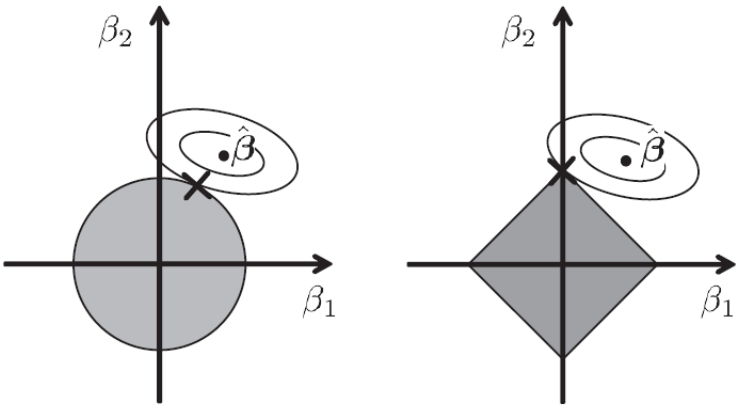
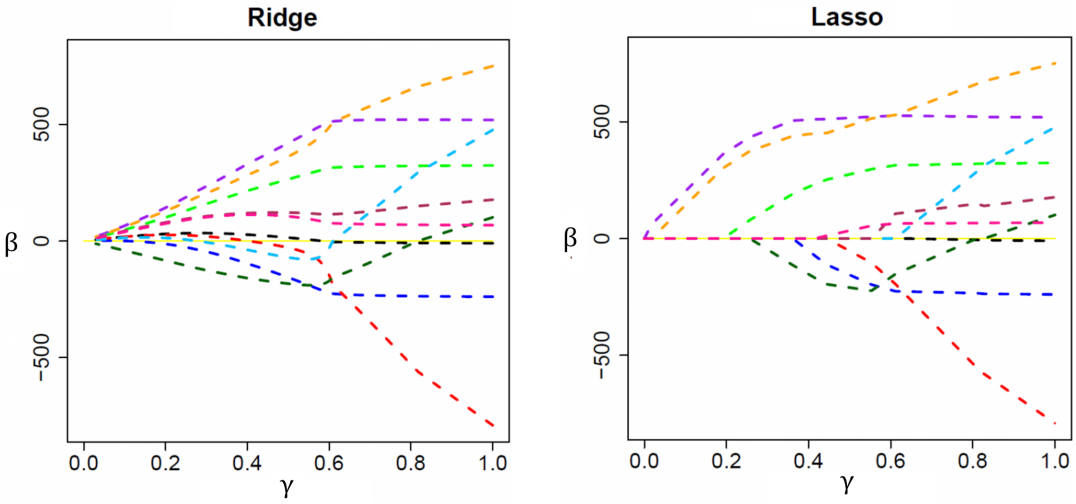


図 2.4: Ridge(左) と LASSO(右) のパラメータ推定の違い
（日本統計学会誌 第 39 巻 第 2 号 (2010) より）



(a) Ridge の変数縮小の様子

(b) LASSO の変数縮小の様子

図 2.5: 変数縮小の様子 [8]

2.2.2　交差検定法

データ解析によって得られた予測モデルは、学習用データに過剰適合せず未知のデータに対しても適応力がある一般化されたモデルである必要がある。本研究では、交差検定法（図 2.6）により、LASSO により得られたモデルの予測性能を定量的に評価する。交差検定法とは、全データを学習用と評価用に分割し、学習用データを用いてモデル構築、評価用データを用いてモデルの妥当性の検証・確認に当てる手法である。交差検定法を用いることにより過剰適合に陥らず一般化されたモデルの獲得が可能となる。今回は交差検定法の結果、平均予測誤差が最小となるモデルを最適モデルとして採用する。

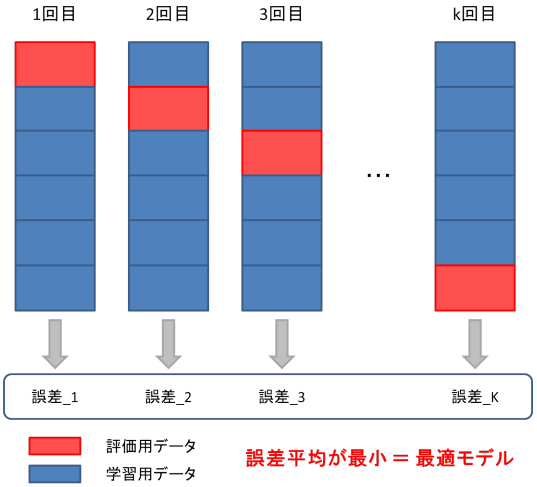


図 2.6: k-分割交差検定法

尚、本実験では、交差検定法として leave-one-out 法を採用した。leave-one-out 法とは、データ群から 1 つだけデータを抜き出しそれを評価用データ、残りの全てのデータを学習用データとして使用し、全データが一回ずつ評価用として使用されるまで検証を繰り返す交差検定法である。

2.3　グラフ化

2.3.1　並列重回帰分析によるグラフ構築

グラフ構造はデータに潜む複雑な構造を可視化することが出来る一般的で高い記述力を持ったデータ形式である。Web、生物系、ビジネスなど実世界の多くの場面でグラフ構造を持ったデータが見受けられるが、データマイニングにおいてもその重要性は認識されており、グラフ構造を扱うためのデータ解析手法の開発が進められている。

LASSO によって得られた線形回帰モデルは、各変数をノードとみなし、回帰係数がゼロでないノード同士をエッジで結ぶことによりグラフ化することが出来る（図 2.7）。Meinshausen と Buhlmann は LASSO を用いて高次元の複雑なデータからスパースなグラフィカルモデルを作成した [11]。このモデルは内部構造の予測モデルであり、かつ、データの持つパラメータ間の関係性を描写するものである。彼らの主張は各変数を対象に LASSO を行い結果をまとめ上げることによって、統計学的に一致性を持つ構造学習が行えるというものであった。本研究においても Meinshausen-Buhlmann の手法にならい、各合金群の融点予測モデルにおいて予測対象を順次入れ替え並列に本解析手法を実行することにより、各合金群に内在する物性関係の全体像を獲得する（図 2.8）。

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

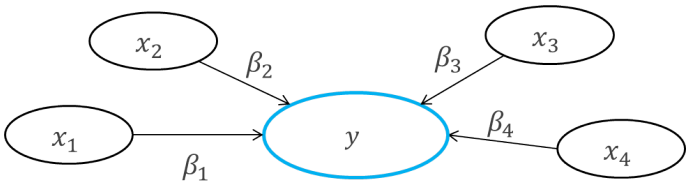


図 2.7: 線形回帰式のグラフ化

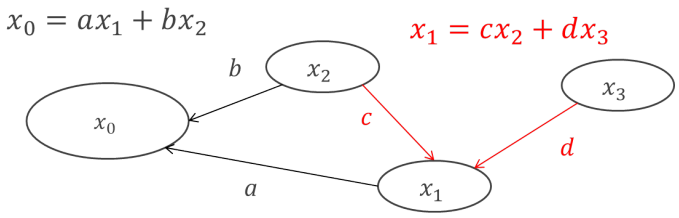


図 2.8: 並列重回帰分析とグラフ化による内部構造の可視化

2.3.2 変数重要度の測定

LASSO により得られた予測モデルの回帰係数は、あくまで予測値を算出する上での値を意味しており、その値の大きさが予測における変数の重要度を意味するものではない。例えばデータの単位を mg から g に変換することにより回帰係数は 1000 分の 1 となるが、その変数が予測において占める重要性は変化しないはずである。そこで、どの変数が予測の上で重要であるか測定するために、以下の手順で変数の重要度の測定を行う。

融点予測モデルにおいて

1. j 番目の説明変数のみを除外し、予測モデルを再構築する
2. 予測誤差を求め、除外した変数の重要度を測る
3. $j+1$ 番目の説明変数のみを除外し、予測モデルを再構築する
4. 以降、全ての変数に対してこの方法で重要度の測定を行う

ここで、2. の変数重要度の測定について詳細を説明する。目的変数（予測対象）と説明変数の間の関係性が弱い場合、その説明変数を除外したモデルであっても予測誤差はあまり変化しないはずである。逆に目的変数と説明変数の間の関係性が強い場合、その目的変数を除外することにより予測誤差は大幅に変化するはずである。従って、変数の重要度の測定に説明変数を除外した際の予測誤差の値を用いることにする。

また、元の予測モデルの $score$ も考慮すべき要因である。$score$ は決定係数（寄与率）と呼ばれるもので、説明変数により予測対象をどの程度説明できるかを表す指標であり、

モデルのデータへの当てはまりの良さの尺度として利用される。$score$ は次の式で与えられる。

$$score = 1 - \frac{\Sigma_{i=1}^{m} \left({y_i}^{obs} - {y_i}^{predict} \right)^2}{\Sigma_{i=1}^{m} \left({y_i}^{obs} - y^{mean} \right)^2}$$

ここで、y^{mean} とは y の平均値である。$score$ は 0 から 1 までの間の値を取り、値が 1 に近いほど正確な予測ができていると言える。

$score$ が小さいということは、そのモデルの予測対象と説明変数の関係性がそもそも弱いことを意味している。従って、仮に予測誤差の値が同じであっても、元の関係性が強い場合と弱い場合を同列に扱う訳にはいかない。以上を踏まえて、変数の重要度を次の式で測定することにする。

$$I_j = score \times \frac{\hat{R}_j}{\Sigma_{i=1}^{m} \hat{R}_i}$$

ここで、$\hat{R}_j$ は説明変数 x_j を除外した時の予測誤差、$\sum_{i=1}^{m} \hat{R}_i$ は全合金データにおける $\hat{R}_j$ の総和である。I_j は 0 から 1 までの間の値を取り、I_j の総和は $score$ に等しい。I_j の値が大きい変数（x_j）ほど予測において重要な変数であると言える。また、I_j は単なる目的変数と説明変数間の相関の大きさではなく、他のすべての関係性が考慮された標準化後の指標となっている。

以上の本研究で使用する手法の実装には、オープンソースのオブジェクト指向スクリプト言語 python を使用した。LASSO 及び交差検定法による物性予測手法は機械学習ライブラリ scikit-learn[9] を、グラフ化にはグラフ作成パッケージの pygraphviz[10] を用いて実装した。

[9] http://scikit-learn.org/stable/index.html
[10] http://pygraphviz.github.io/

第3章　結果と考察

本章ではまず2章で述べた二元合金データベース及び物性予測モデル構築手法を用いて二元合金の融点予測を行い結果を評価する。続いて、融点予測モデルを拡張し、物性関係の全体像をグラフ化する。最後に、そのグラフを以って合金の物理の理解を行う。

3.1　データの分類

融点予測に先立ち、まずは2章で述べた103種の二元合金データを以下の4つの合金群に分類する。

1. アルカリ金属合金群（15）
2. アルカリ土類金属合金群 (15)
3. 遷移金属合金群 (19)
4. 希土類金属合金群 (54)

分類のルールとして、二元合金ABについて構成原子Aがそれぞれアルカリ金属・アルカリ土類金属・遷移金属・希土類金属の該当する合金群に振り分けるものとする。このルールに従って分配を行った結果、各合金群のデータ数はカッコ内の数値のとおりとなった。これらの合金群に対して本解析手法を適用し、融点予測モデルを構築する。

3.2　融点予測結果と考察

3.2.1　融点予測結果

各合金群の融点予測結果を図 3.1 と図 3.2 および表 3.1 から 3.4 に示す。図は横軸を融点の実測値、縦軸を融点の予測値としており、予測が正確である場合は対角線上にプロットが集中する。表は融点予測モデルに組み込まれた説明変数（物性）とその回帰係数および切片である。

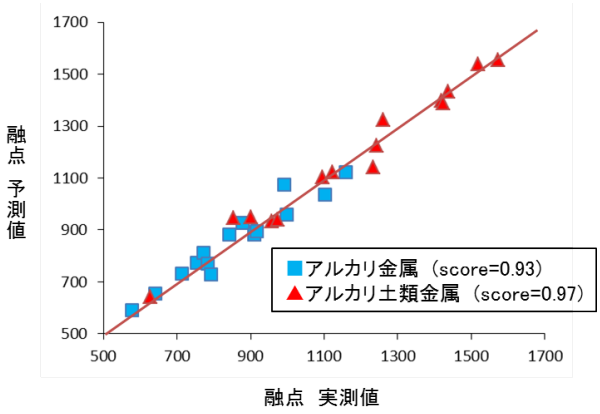


図 3.1: アルカリ金属・アルカリ土類金属合金群の融点予測結果

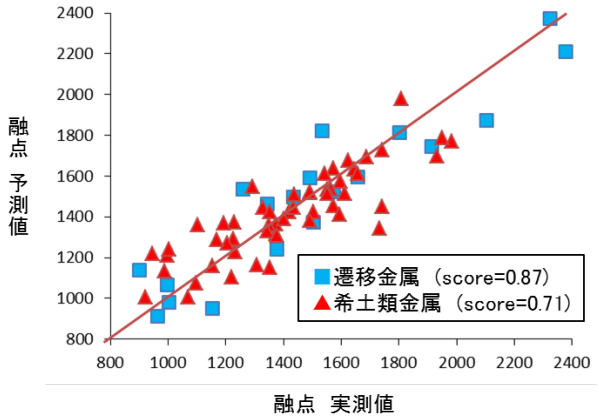


図 3.2: 遷移金属・希土類金属合金群の融点予測結果

表 3.1: アルカリ金属合金群の融点予測モデル

物性	回帰係数・切片の値
$Vapo(A)$	8.59
$BP(B)$	-0.27
$IP(B)$	211.65
$AN(B)$	-13.77
Gap_{BCl}	106.55
$(Dis_{AB})^2$	160.67
Gap_{BAB}	-78.35
$intercept$	-1337.44

表 3.2: アルカリ土類金属合金群の融点予測モデル

物性	回帰係数・切片の値
$VE(B)$	-714.93
$BP(B)$	-0.10
$Vapo(B)$	2.70
$AN(B)$	7.40
Gap_{BCl}	-135.24
$Freq_{AB}$	-0.74
$MC_{ABA}(B)$	-1570.7
$intercept$	1416.12

表 3.3: 遷移金属合金群の融点予測モデル

物性	回帰係数・切片の値
$Vapo(A)$	1.83
$BP(B)$	0.38
$(Dis_{AB})^2$	108.37
$intercept$	-949.54

表 3.4: 希土類金属合金群の融点予測モデル

物性	回帰係数・切片の値
$VE(A)$	41.46
$BP(A)$	0.19
$AN(A)$	-10.59
$IP(B)$	107.49
$Vapo(B)$	1.84
$AN(B)$	6.00
Gap_{ACl}	-267.52
Gap_{BCl}	-126.32
$MC_{BAB}(B)$	-360.79
$intercept$	-791.82

3.2.2　結果の考察

融点予測モデルを構築した結果、アルカリ金属合金群、アルカリ土類金属はそれぞれ *score* の値が 0.93、0.97 となり、高い予測性能を持つモデルを構築することが出来た。しかし遷移金属合金群では 0.87 とやや値を落とし、希土類金属合金群に至っては 0.71 となった。

この理由として、2.1 節で作成した二元合金データベースに合金群の振る舞いを記述可能な変数が不足していたことが考えられる。本解析手法により得られる予測モデルは変数の線形結合形をしているため、準備した変数が不十分であれば得られる予測モデルも正確になることが出来ない。遷移金属、希土類金属の予測が芳しくなかったことから、d 電子系の振る舞いを記述する物性を変数に加える事により、モデルの予測精度が向上することが期待される。また、今回は 2 つ、及び 3 つの金属原子からなる分子モデルの量子計算データのみを使用したが、このようなシンプルな構造だけではなく、もっと多くの原子を扱った複雑な系モデルのデータを使用することで記述力が増し、予測精度が向上すると考えられる。

また、希土類金属合金群については、サンプル数が 54 個と他の合金群に比べて 2 倍以上のデータ量となっていることも予測精度低迷の理由として考えられる。1.1.3 項で述べたとおりデータ量増加に伴い系が複雑化するため、希土類合金群は他の 3 合金群に比べ複雑性が高い状態にあったと推測される。従って、「二元合金 AB の構成原子 A に着目して振り分ける」という今回の分類ルールでは不十分であり、更に何らかの条件のもと分合金の分類を行う必要性があったと考えられる。

その一方で、今回のような 2 つないし 3 つの原子からなるシンプルな分子モデルのデータから、これだけの精度を持つ予測モデルを構築出来たのは特筆すべき事項である。従って、この結果を以って本解析手法を用いることにより、正確な物性予測モデルを構築することが可能であると結論付けることが出来る。

3.3　グラフ化の結果と考察

3.3.1　グラフ化の結果

先述の 4 つの合金群に対し並列重回帰分析を行い、得られた関係性をグラフ化した結果、図 3.3 から図 3.6 のようになった。このグラフは物性をノード、相関関係の有無をエッジで表した有向グラフであり、矢印の根元のノードの物性が矢印の先のノードの物性に寄与することを表している。エッジの色は相関関係が性の関係の場合は赤色、負の関係の場合は青色としている。また、グラフ中の文字 R は変数の重要度、A はその変数によって予測対象物性の値が変位する範囲であり、次の式で算出している。

$$A_j = I_j \times \beta^j \times x_j \text{の取る値の範囲}$$

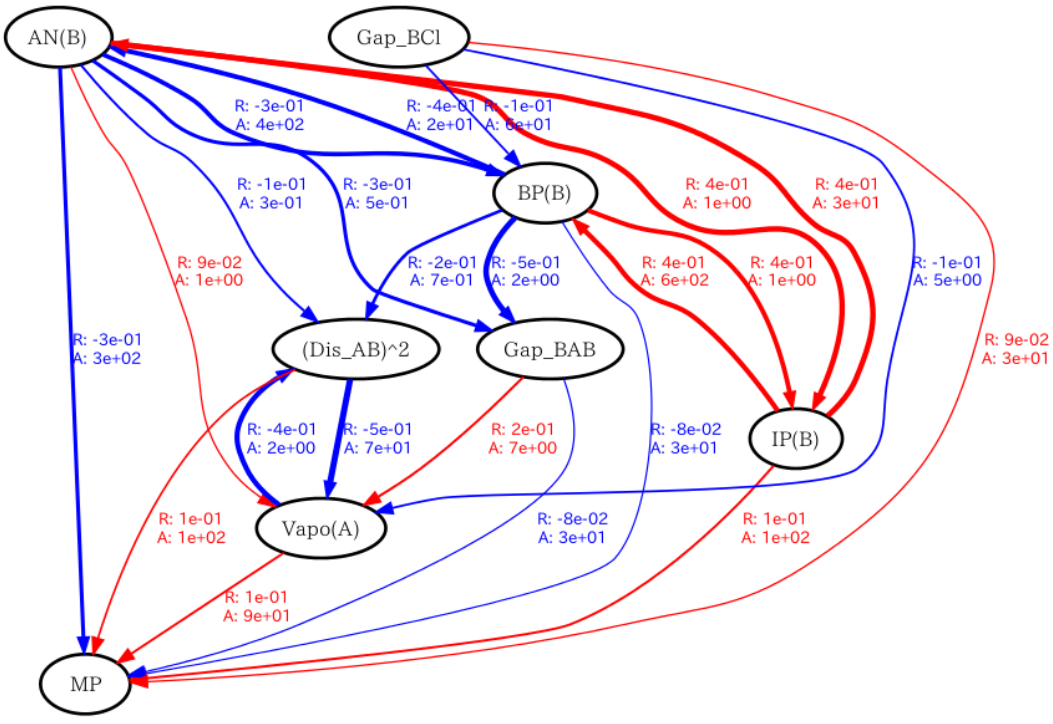


図 3.3: アルカリ金属合金群のグラフ

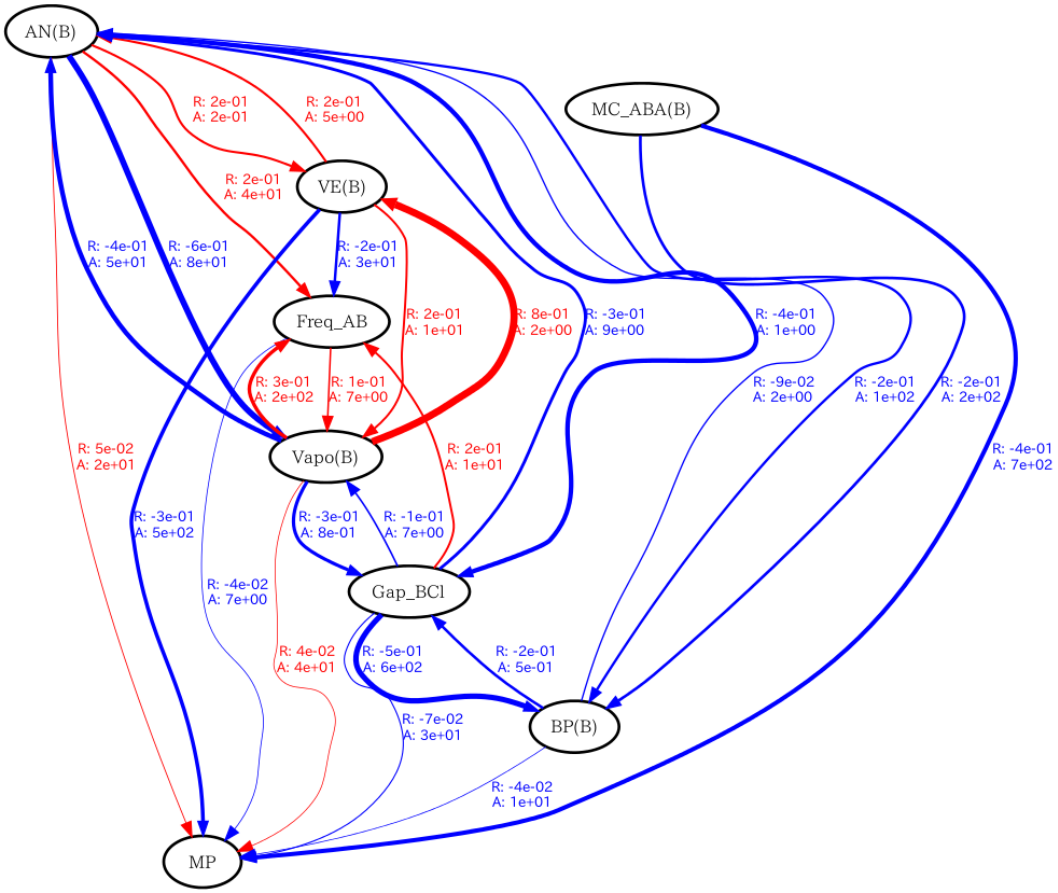


図 3.4: アルカリ土類金属合金群のグラフ

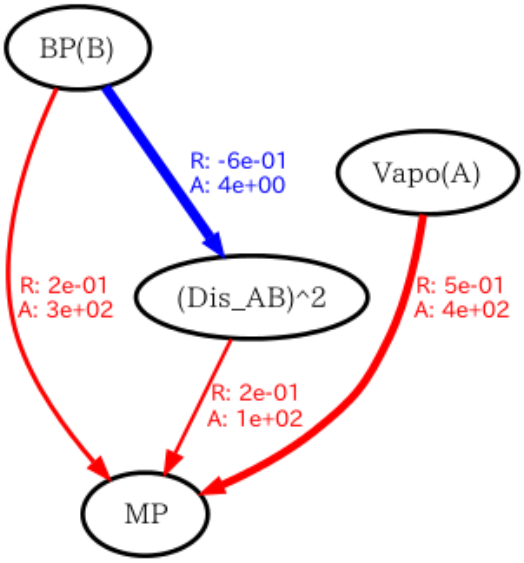


図 3.5: 遷移金属合金群のグラフ

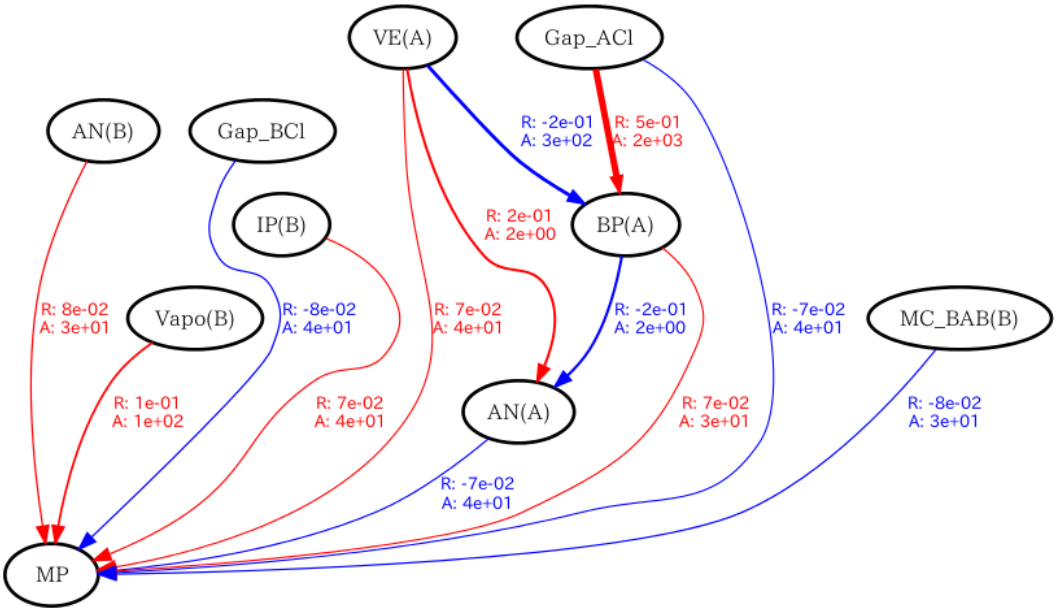


図 3.6: 希土類金属合金群のグラフ

3.3.2 結果の考察

グラフの全体的な形としては、アルカリ金属合金群とアルカリ土類金属合金群のグラフは変数間の関係性がよく現れているが、遷移金属と希土類金属は変数間の関わり合いが少なく、融点に向かう関係性のみが多く現れている。ここから、アルカリ金属及びアルカリ土類金属合金群については、合金群の物理を記述する上で必要な物性を取り込むことが出来たが、遷移金属及び希土類金属合金群は充分に取り込めていなかったことが分かる。それを顕著に表しているのが遷移金属合金群の結果である。遷移金属合金群の融点予測モデルは $score$ の値的には 0.87 と悪くなかったが、グラフを見ると変数間の関係性は全く現れておらず、また、融点予測も気化熱（$Vapo(A)$）と沸点（$BP(B)$）に大きく依存している。気化熱や沸点が融点と関係がありそうなことは直感的にも明らかであり、有用な知識であるとは言えない。これは本解析手法が交差検定法を採用しており、全二元合金データを平均的に良く予測出来るモデルを最適モデルとして選択する性質を持つために発生した現象であると考えられる。つまり、合金群の振る舞いを記述するのに充分な物性が用意されなかったが、融点の予測に関してのみ充分に記述出来る変数が存在したことにより、このようなグラフが出来上がってしまったのである。このように、グラフ化することにより物性間の関係性と変数の重要度が分かるだけでなく、融点予測のような一つの物性に注目した場合は気づくことが出来なかった予測モデルの出来の良し悪しも判断可能であることが分かった。

続いて、比較的物性間の関係性を構築することが出来たアルカリ金属・アルカリ土類金属合金群のグラフについて、物理的観点から評価を行う。グラフから、アルカリ金属合金群ではイオン化エネルギー（$IP(B)$）や HOMOLUMO ギャップ（Gap_{BAB}, Gap_{BCl}）が、アルカリ土類金属合金群では価電子数（$VE(B)$）やマリケン電荷（$MC_{ABA}(B)$）、HOMOLUMO ギャップ（Gap_{BCl}）が融点予測の上で重要な変数として確認できることから、両合金群共に融点には電荷移動や反応性が深く関わっていることが読み取れる。ここで、物理的には物質が融解するということは、元素間の結合が切れることを意味しており、また、アルカリ金属・アルカリ土類金属は最外殻の s 軌道の電子だけが結合に関与することで知られている。従って融点予測に関するグラフにおいて電荷や結合に関する物性が現れるのは、合理的な結果であると判断できる。

また、アルカリ金属合金群は構成要素 B の原子番号（$AN(B)$）が重要である一方、アルカリ土類金属合金群は、構成要素 B の価電子数（$VE(B)$）が重要であるという結果が得られた。ここで、構成元素 B に注目したところ、アルカリ金属合金群で使用されてい

た構成元素 B は 11 族、13 族、14 族、15 族の 4 つの族に属していたのに対し、アルカリ土類金属は 12 族、13 族、14 族の 3 つの族のものとなっていた。つまり、アルカリ土類金属の方が価電子数のバリエーションが少ないため、結果として強く現れてしまった可能性がある。従って、バリエーションを揃えて分析することにより正当な評価を行うことが必要である。

一方、この結果は妥当である可能性もある。アルカリ金属は反応性が非常に強いため、相手の価電子にあまり依存せずそれ以上に質量が重要視される、そして、アルカリ土類金属は 2 価の価電子を持つので、アルカリ金属と比べると結合相手の価数に対して多様な電子状態を取ることが出来るためこのような結果が現れた可能性がある。

第4章　まとめと今後の課題

4.1　まとめ

本研究は帰納（データマイニング）と演繹（量子計算）の融合した物性予測手法の提案を行ってきた。提案手法の有効性を示すため、具体例として二元合金の融点予測および予測モデルにおける各物性間の関係性の説明の 2 つの課題に取り組んだ。

まず、融点予測においては、二元合金データを実験/基礎物性データと、２つ及び 3 つの原子からなる分子モデルの量子計算データを用いて表現し、データベースを作成、そのデータベースに対して LASSO を実行することにより融点予測の線形回帰モデルを獲得した。4 つの合金群について本手法を適用し得られた結果を見ると、アルカリ金属・アルカリ土類金属合金群は *score*=0.9 以上のモデルを得られたが、遷移金属・希土類金属のモデルはそれぞれ 0.87、0.71 に留まった。これは今回実験で準備した物性データだけでは遷移金属・希土類金属の振る舞いを記述しきれていないことを意味しており、例えば d 電子系の振る舞いを記述するような物性を追加することによりさらなる予測精度の向上が見込める。その一方で、このようなシンプルなモデルのデータから正確な融点予測が行えたことは特筆すべき点である。

続いて、融点予測モデルを拡張し、モデルに含まれる物性の間の関係性をグラフ化した。具体的には各融点予測モデルに含まれる各変数を予測対象に LASSO を行い、結果をまとめ上げることによりグラフ化を行った。また、*score* と変数除外時の誤差を使用して予測における変数重要度を測定した。グラフからは、アルカリ合金・アルカリ土類金属合金群は共に電荷移動が融点予測において重要なファクターであることや、構成要素 B について、アルカリ金属は質量が重要である一方、アルカリ土類金属は価電子数が重要であるというような性質の違いをもつ可能性があることを確認出来た。一方、遷移金属・希土類金属は充分な変数を用意して予測が行えていなかったことがグラフからも判断できる結果となった。このように変数重要度や変数間の関係性など、単に物性の予測を行うだけでは得難い情報も、グラフ化により獲得できることを確認した。

以上の結果を以って、本提案手法の有効性を示すことが出来た言える。

4.2　今後の課題

今回は遷移金属・希土類金属については良い結果を得ることが出来なかった。従って予測を行う上で必要であると思われる変数、例えばｄ電子系の振る舞いを記述できる変数を追加し、遷移金属・希土類金属群に関しての知識を獲得できるかどうかを検証する必要がある。また、ｄ電子系の変数を追加することで性質を描写できるようになるとは限らない。従って、予測に関わる重要な変数を効率的に発見する方法論の開発も必要であると考えられる。また、今回はグラフ作成までは行ったが、材料設計に応用するまでには至らなかった。最適化、検索、クラスタリングなどのグラフアルゴリズムを用いることにより、本研究で得られたグラフを活用した効率的材料設計法を考案することが今後の最重要課題である。

謝辞

本論文を執筆及び日々の研究を行う上で、私は多くの方々にお世話になりました。

主指導教員の Dam Hieu Chi 先生には大変お世話になりました。本研究を行うにあたり、統計額・物理学・プログラミング…その他諸々スキルや知識が必要でしたが、それらの全てに不足していた私がどうにか完走出来たのも先生のご指導のお陰に他なりません。本当にありがとうございます。また、一度博士後期過程に進学を表明しながら、一身上の都合により急遽取りやめご迷惑をお掛けしたこと、深くお詫び申し上げます。

杉山 歩先生には本論文のみならず、分子科学会のポスター発表や学内研究ユニットにおける発表など、大変お世話になりました。先生の客観的なアドバイスのお陰で内容を大分整理することが出来ました。また、研究のみならず日々の生活においても白山の素晴らしい自然、美味しいお店を教えていただきましたこと、非常に感謝しております。

水上 卓先生も、私の調子が良くない時に気遣ってくださったり、昼食に何度かお誘い頂きまして大変お世話になりました。その際に先生と話した事柄や美味しい料理のお陰でリフレッシュし、新たな気持で研究活動を行うことが出来ました。

研究室メンバー、知識科学研究科 同期各位にも感謝致します。私が楽しく大学院の生活を送ることが出来たのは皆様のお陰です。正直に言いますと、社会人経由で入学した私は入学前に、「年齢が離れているので周りと上手くやっていけるか？お互いやりにくいところがあるのではないだろうか？」と多少心配していました。しかし、そんな心配は全く無用でした。むしろ年上の人間に対してこんな風で、社会でやっていけるんだろうか？と心配になるほどでした。

最後に、大学院での研究生活を支援してくれた家族、そして退職後も切磋琢磨する存在であってくれた前職の仲間たちに深く感謝します。

参考文献

[**1**] http://www.whitehouse.gov/mgi

[**2**] http://www.jst.go.jp/crds/pdf/2013/SP/CRDS-FY2013-SP-01.pdf

[**3**] http://www.jst.go.jp/crds/pdf/2013/WR/CRDS-FY2013-WR-03.pdf

[**4**] 陳迎, 金田保則, 川口福太郎, 岩田修一, P. Villars, 材料設計のためのデータシステム－逆問題への適用 (2003)

[**5**] Yousef Saad, Da Gao, Thanh Ngo, Scotty Bobbitt, James R. Chelikowsky, Wanda Andreoni, PHYSICAL REVIEW B 85 104104 (2012)

[**6**] R. Tibshirani, J. R. Statist. Soc. B 58, 267 (1996).

[**7**] 日本統計学会誌 第 39 巻 第 2 号 (2010), pp.211-242.

[**8**] https://onlinecourses.science.psu.edu/stat857/node/158(2014.2.1 アクセス)

[**9**] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Annals of Statistics 32, 409 (2004).

[**10**] 鹿島 久嗣, グラフとネットワークの構造データマイニング, 電子情報通信学会誌 93(9), 797-802 (2010).

[**11**] Meinshausen, N and Buhlmann, P, Ann. Statist., 34, 14361462(2006).

[**12**] 井手剛, 潜在的グラフ構造からの異常検知, Technical Report of the 1st Workshop on Latent Dynamics(2010)

付 録A　二元合金データの持つ属性一覧

実験/基礎データ（13 個）

表記	モデル	意味
MP	A-B	二元合金の融点
$VE(A)$	A-B	A の価電子
$EN(A)$	A-B	A の電気陰性度
$BP(A)$	A-B	A の沸点
$IP(A)$	A-B	A の第一イオン化エネルギー
$Vapo(A)$	A-B	A の気化熱
$AN(A)$	A-B	A の原子番号
$VE(B)$	A-B	B の価電子数
$EN(B)$	A-B	B の電気陰性度
$BP(B)$	A-B	B の沸点
$IP(B)$	A-B	B の第一イオン化エネルギー
$Vapo(B)$	A-B	B の気化熱
$AN(B)$	A-B	B の原子番号

量子計算データ（17 個）

表記	モデル	意味
Gap_{ACl}	A-Cl	ACl（塩化物）の HOMOLUMO ギャップ
Gap_{BCl}	B-Cl	BCl（塩化物）の HOMOLUMO ギャップ
Dis_{AB}	A-B	結合距離
BE_{AB}	A-B	結合エネルギー
$HC_{AB}(A)$	A-B	A のヒルシュフェルト電荷
$Freq_{AB}$	A-B	振動周波数
$(HC_{AB}(A))^2$	A-B	A のヒルシュフェルト電荷の 2 乗
$(Dis_{AB})^2$	A-B	結合距離の 2 乗
CF_{AB}	A-B	クーロン相互作用
Dis_{ABA}	A-B-A	結合距離
Gap_{ABA}	A-B-A	HOMOLUMO ギャップ
$MC_{ABA}(A)$	A-B-A	A のマリケン電荷
$MC_{ABA}(B)$	A-B-A	B のマリケン電荷
Dis_{BAB}	B-A-B	結合距離
Gap_{BAB}	B-A-B	HOMOLUMO ギャップ
$MC_{BAB}(A)$	B-A-B	A のマリケン電荷
$MC_{BAB}(B)$	B-A-B	B のマリケン電荷

付 録B　二元合金データー覧

全 103 種、アルファベット順

BaCd	ErCu	LaAg	NaIn	SmTl	YGa
BaGe	ErGa	LaAu	NaPb	SrCd	YGa
BaHg	ErIn	LaCd	NaTl	SrGe	YIn
BaPb	ErNi	LaHg	NdAg	SrSi	YNi
CaCd	EuAg	LaNi	NdPt	TbAg	YNi
CaGe	EuAu	LaTl	NdSi	TiCu	YZn
CaHg	EuIn	LiAg	NdTl	TiPt	YZn
CaSi	EuPb	LiAl	PrAg	TmGa	ZnAg
CaSn	GdAg	LiAu	PrAu	YAg	ZnAu
CaTl	GdCu	LiBi	PrGa	YAg	ZnCu
CdAg	GdRh	LiGa	PrNi	YbAg	ZrIr
CdAu	GdTl	LiIn	RbAu	YbAu	ZrNi
CeAu	HfCo	LiPb	ScAg	YbCd	ZrPt
CeZn	HfNi	LiTl	ScAg	YbGa	
DyAu	HoAg	LuGa	ScAl	YbNi	
DyCu	HoGa	MgHg	ScAl	YbPd	
DyIn	KPb	MgTl	SmAg	YbTl	
ErAu	KSn	NaBi	SmNi	YbZn	