修　士　論　文

東アジアにおける酸性･酸化性物質の動態解明へのクラスタリング手法の適用

指導教官　　ホー ツー バオ 教授

北陸先端科学技術大学院大学
知識科学研究科知識システム基礎学専攻

850018　小山内　尚

審査委員：　ホー ツー バオ　教授（主査）
石崎　雅人　助教授
中森　義輝　教授

2000 年 2 月

目　次

図　目　次

表　目　次

第 1 章
はじめに

地球環境問題とはそもそも人間活動の拡大に起因する。環境を無視した開発に対して早い時期から「自然は人間に仕返しをする」という警告が出されていた。

大気中の炭酸ガス濃度の増加による地球温暖化、フロン化合物によるオゾン層破壊は単なる特定の地域問題から地球環境問題へと転換する重大なきっかけとなった。大気中の炭酸ガス濃度の増加については大気のみではなく、海洋、陸地との関係も無視することはできない。これらは、不確かなところが多く残り将来予測の困難さをもたらしている。

酸性降下物は本来地域に限定される汚染物質であったが、比較的早くから気流に乗った越境が問題となり、地球規模の汚染として注目されるようになった。酸性雨被害は北欧やカナダのように顕著にはまだ出ていないが、大陸における発生量の今後の増大がもたらす結果を予測して対策を立てねばならない。

1950年には森林が地球の陸地の約30％を覆っており、その内約半分は熱帯林だった。しかし2000年には熱帯林は1950年当時の5％程度にまで減少するであろうと言われている。1975年頃までは熱帯雨林の需要は主として材木とパルプだったが、現在は50％が薪として使われている。それでも薪は不足しており、藁や牛糞など４億t以上が燃料に使われ大地へ還元されていない。農地に転換しても土壌中の有機物、栄養塩類の枯渇は癒しようがなくなっている。アジア、アフリカの人口増加がこの傾向をさらに大きくすることは確実である。南アメリカにおいては広大な地域が無樹の大平原として牧畜に用いられている。非常に低価格の牛肉がここから供給されているが、地球における森林の価値を計る尺度が、いまだに材木として利用されること以外に存在しないためである。

地球環境問題で確かなことは人口の増加が生物学的論理で抑制されるということ、つまり食糧、絶対生活空間の限界による抑制がどこかで働くということである。持続可能な開発がどの程度まで行えるか。そのためには資源の循環活用が必須であるが自然生態系抜きの技術依存型だけで出来るとは考えられない。

海洋汚染も森林伐採もともに資源の枯渇をもたらす重要な要因である。そしてまだ記録さえされていない多数の生物種を絶滅させている。ある生物種の生存はその属する生態系を維持する上で欠くことができないものである可能性が高い。さらに極めて重要な遺伝子資源を失うことになることが考えられる。医療に使用される薬用動植物は現在も数多く、アメリカにおいてさえ全体の25％は高等植物を起源としているものであるという。抗マラリア剤としてのキニーネとその化学構造を改変した合成薬剤のクロロキンが広く使われ、最近では中国の蓬の一種から抽出したアルテミシンが有望視されている。また土壌細菌から見つけられた副作用が少ない抗フィラリア剤のアイベルメクチンが使われるようになったのもごく最近のことである。

地球環境問題においては、地球環境変化の影響の程度についても明らかではない。人に対する直接的な影響もさることながら、地球規模の変化は多岐に渡りしかも連鎖反応を引き起こしていく事が予測されるが、未だ十分納得のいく研究はなされていない。そして、オゾン層の破壊が紫外線の増加をもたらし、人間や生態系にも影響をあたえることも十分に考えられる。また、温暖化による気候変動の結果がどうなるのかも予測しがたいところがある。

地球環境は、ある限界を越えて問題が顕在化したときにはもう手の打ちようがないという面を持つ。蓄積された廃棄物、破壊された環境規模の膨大さ故に、それを修復する手だてを人間は持っていない。突き放した言い方をすれば、あらゆる産業活動をただちに停止し、大自然の大きな浄化作用に身を任せるしか解決策はないのだろう。たとえ、最新技術によって人間に授けられる解決策を用いても、それは新たな問題を呼び起こしてしまう。そして、科学により地球環境の悪化を食い止める切り札を誰一人として持っていないのである。

しかし、技術の進歩には目覚しいものがある。コンピュータの性能の向上とともに計測データの電子化、データベース化が著しく進み、膨大なデータの集積が可能となってきた。このような電子化の進展によって集積された膨大なデータは人の処理能力をはるかに越え、計算機による有効な使用方法の確立の必要性が指摘される

ようになった。今もなお蓄積されつつある膨大なデータベースからの有効な知識の抽出に対して、データマイニング及び KDD の手法の有効性が期待されている。

このような時代の流れから、データマイニングを行うこと地球環境問題を改善する方法を見出すことが考えられる。ひとくちに地球環境問題といっても地球の温暖化・オゾン層破壊問題・酸性雨など様々な問題があり、しかも全てを統合したデータベースは存在しない。そこで、本研究では、国立環境研究所によって収集･蓄積され、研究者へ提供されている航空機調査におけるオゾン、窒素酸化物、および二酸化硫黄の観測データを結合し用いることにした。これにともない、地球規模の問題から東アジアの問題へと焦点を絞った。データ量は一般的なデータマイニングという観点からは小さいが、より具体的にデータマイニング手法の 1 つであるクラスタリング手法を用いることとした。

東アジア地域は、人口の増大と急速な工業活動の発展により、窒素酸化物と二酸化硫黄の人為的排出が最も多い地域の 1 つである[Rodhe 89]。加えて、21 世紀には、この地域の窒素酸化物と二酸化硫黄の排出量が、世界最大になると予測されている[Galloway 89]。そこで、東アジア地域における現在の大気の汚染状態の分析を行い、さらに、未来の大気の汚染状態をコンピュータモデルなどを用いて予測し、対策を立てることは非常に重要となってくる。対流圏バックグラウンドのオゾンは大気の酸化能を左右する重要な化合物である。同時にまた、温室効果ガスとして、地球温暖化の観点からも非常に重要である。近年、北半球における対流圏オゾンの増大が欧米で報告され、注目を集めている。対流圏におけるオゾンの前駆体となるのは二酸化窒素のみであり、窒素も含めたいわゆる窒素酸化物の人為的放出の増大が、このようなオゾン濃度の増加をもたらしているという指摘がある。このようにオゾンおよび窒素酸化物は、様々な角度からみて対流圏大気の化学を支配している重要な因子であり、その立体的な濃度分布を知ることは、対流圏大気の変動を考える上で欠かすことのできないファクターである。一方、二酸化硫黄は酸性雨原因物質の硫酸の前駆体として重要な役割を果たしている。この二酸化硫黄は、中国や韓国を含む東アジア地域から大量に移流してくると考えられている[地球環境センター 96]。

従って、本研究では、クラスタリング手法を適用し、東アジアにおける酸性･酸化性物質の動態解明に関する有効な知識、特に気候モデルと予測に対し、深い影響力をもっているレジムの発見を目的とする。

第 2 章
地球環境問題とデータマイニング

2.1 本研究で扱うデータについて

本研究では、地球環境研究センターによる地球環境データを研究者に提供する「地球環境研究支援」という業務の一環として CD-ROM として刊行された「'97 IGAC/APARE/PEACAMPOT 航空機・地上観測データ('91～'95 集成版)」を用いる。これは、平成 3 年度より平成 7 年度にかけて、環境庁地球環境研究総合推進費 酸性雨分野課題「東アジアにおける酸性・酸化性物質の動態解明に関する研究」によって行われた航空機観測及び集中地上観測のデータを集成したものであり、データは数値データ利用者のために作成されているので非常に扱いやすいものとなっている。

そして、実際には「'97 IGAC/APARE/PEACAMPOT 航空機・地上観測データ('91～'95 集成版)」の"フライトデータ"とそれに対応する"IGAC/APARE/PEACAMPOT 航空機調査におけるオゾン（O_3)、窒素酸化物(NO_x)、および二酸化硫黄（SO_2）の観測"の 5 年分である。

本研究で扱う国立環境研究所地球環境研究グループ酸性雨研究チームによるデータの主要な成果としては、酸性雨原因物質の長距離輸送とその輸送パターン、特に輸送経路による汚染物質濃度の違いである。平成 8-10 年度の観測結果からは、中国中南部からの輸送パターンが明らかとなり、国立環境研究所で進めてきたモデルの結果と非常によく一致する結果が得られている[Hatayama 97]。

2.2 データマイニング

1995年にモントリオールで行われた、最初のKDD国際会議において、KDDとはデータから知識を抽出するプロセス全体のことであると提案された。この提案において知識というものは、データの要素間の関係とそのパターンを意味する。さらに、データマイニングという用語は、KDDの過程の中の発見の段階だけに使用することが提案された。KDDの定義は、「妥当性・新規性・潜在的有用性・最終的に理解可能性のあるパターンをデータから同定するための自明でないプロセス」である。データマイニングの定義は、「KDDプロセスの１つのステップであり、データ分析とデータ発見のアルゴリズムを適用して、計算効率を考慮しつつ、データ中の特定のパターンを数え上げ、抽出するもの」である。

KDDは全く新しい技術ではなく、様々な研究を融合した領域である。機械学習、統計学、データベース技術、エキスパートシステム、そしてデータ可視化技術のすべてがこれに貢献している（図2.1)。

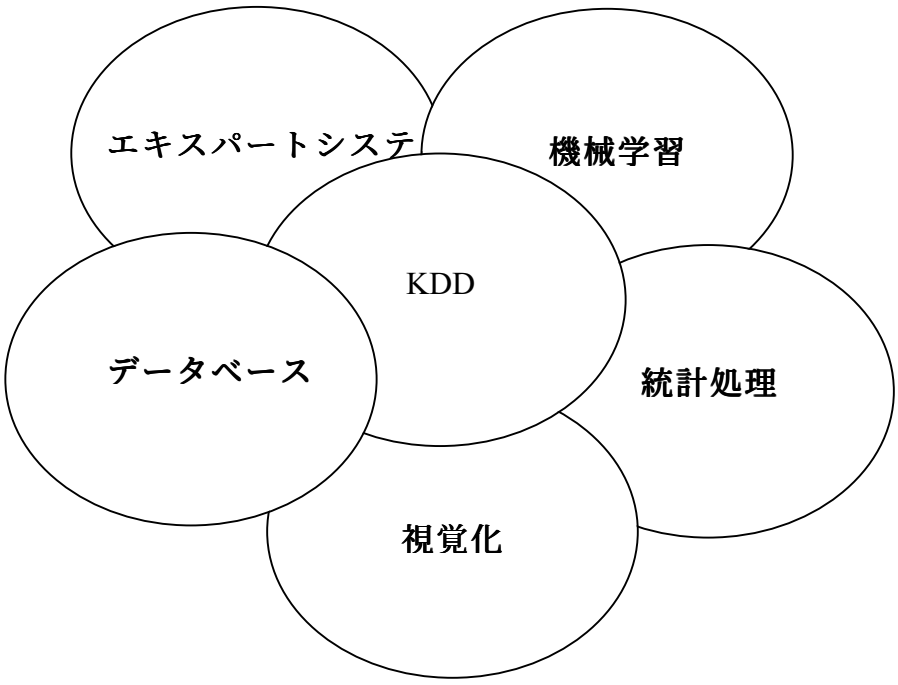


図 2.1 関連分野[エイドリアン 98]

一般に、人々がデータマイニングについて語るとき、主に実際のマイニングおよび発見に焦点をあわせている。このアイディアは、直感的に分かりやすそうで魅力的に思われる。しかし、データマイニングは、前にも述べたようにプロセス全体のほんの 1 ステップにしか過ぎないものであり、図 2.2 に示すように、KDD は複数のステップからなる反復プロセスである。

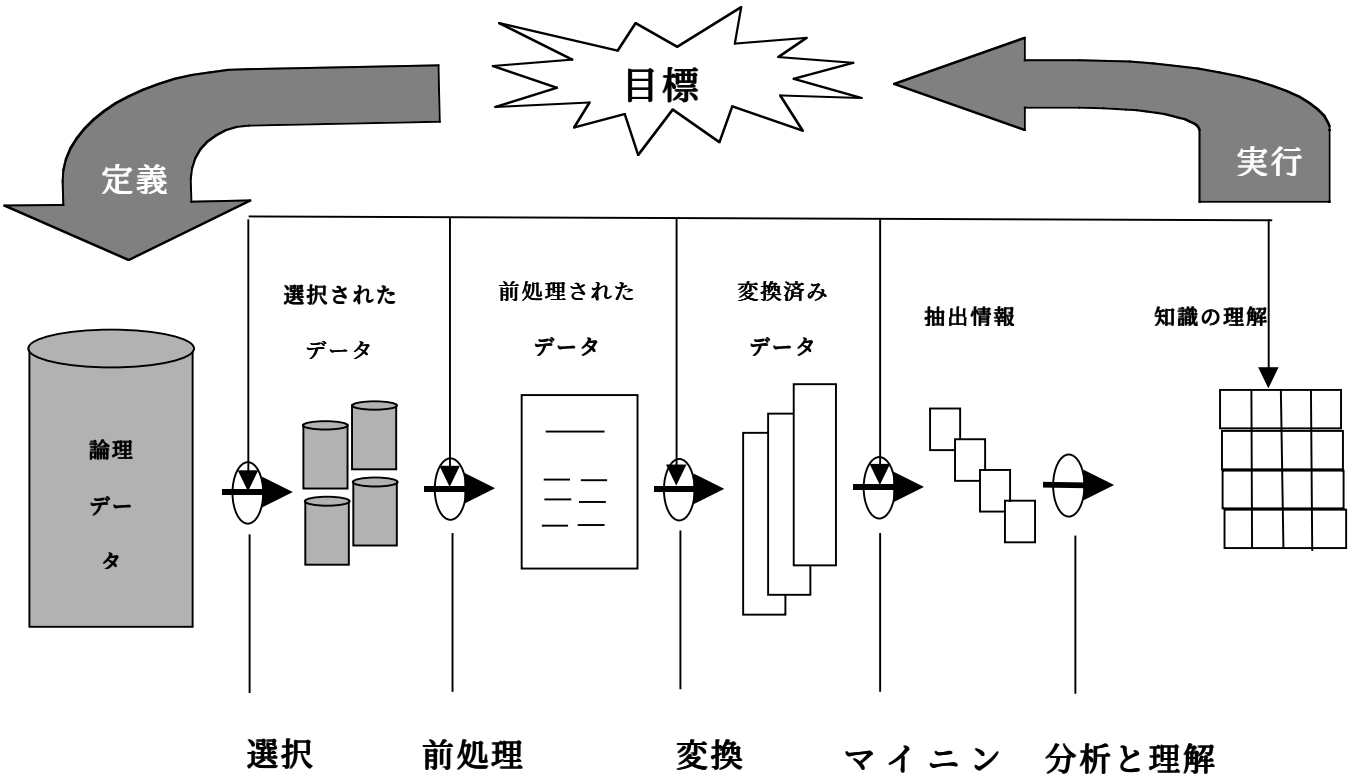


図 2.2 KDD プロセス[キャベナ 99]

地球環境の改善という目標を達成するために、本研究では KDD プロセスを行う。この目標は、最初のプロジェクトを立てる基礎となるものであり、最終結果を評価するための判定基準である。それゆえ、その目標はプロセスの中の多くのステップ全体にわたって、チームを絶えず導くものでなければならない。各ステップは、図 2.2 で示されている順序で行われるが、プロセスはかなり反復的なものとなり、おそらく 1 つ以上のステップが何回も繰り返され、プロセスは自立走行式と呼ばれるものとはかけ離れている。最近のテクノロジーの進歩にもかかわらず、いまだデータマイニングは非常に大きな労力を要する作業である。

大きな労力を要するとは言っても、プロセスのすべてのステップで、普通に費やされる時間や作業量が同じわけではない。図 2.3 は、プロセスの概略ステップとそれぞ
れのステップで、通常必要とされる相対的な作業量を示している。見てのとおり、時間の 60%はマイニングのためのデータ準備に費やされており、このことはクリーンで適切なデータに大きく依存していることを顕著に示している。通常、実際のマイニング・ステップは、作業量全体のおよそ 10%である。

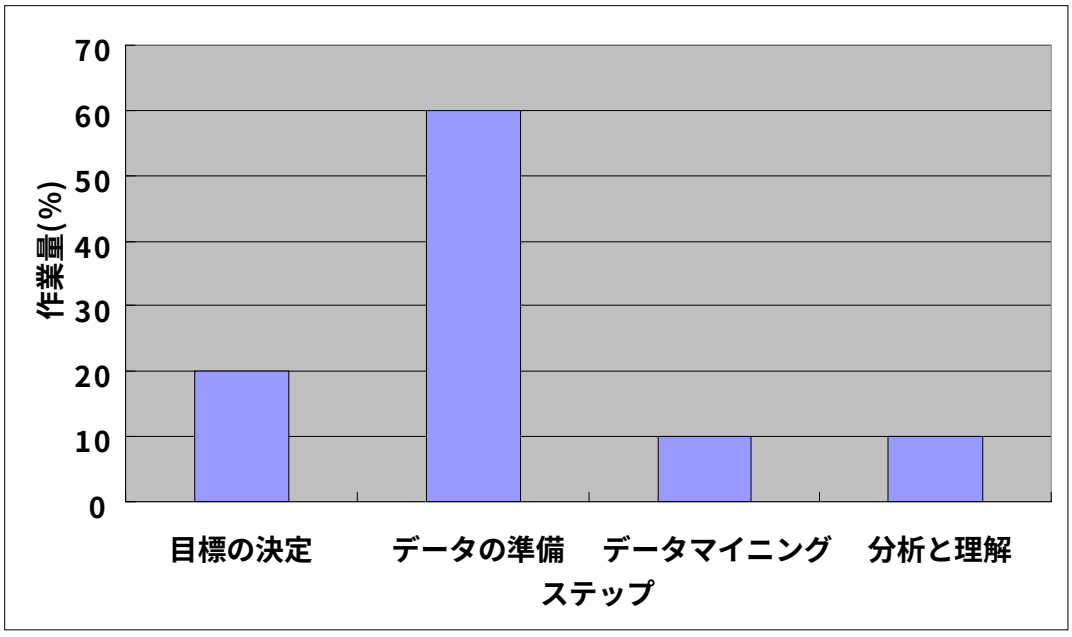


図 2.3 各 KDD プロセス･ステップに必要な作業量[キャベナ 99]

各ステップの概略を以下に説明しよう。

ステップ 1：目標の決定

問題または課題を明確に定義する。これは、どんなデータマイニング･プロジェクトでも不可欠な要素である。直感的に分かり簡単そうに思えるかもしれないが、実際には決してそうではない。

ステップ 2：データの準備

ステップ 2.1：データ選択

情報のすべての内部ソースと外部ソースを識別し、どのサブセットのデータがデータマイニングに必要かを選択する。

ステップ 2.2：データの前処理

分析を容易にするため、また実行を可能にすると同時に、実行する価値のあるマイニング操作の種類を決定するため、データの品質を調べる。

ステップ 2.3：データの変換

データを分析用に適した形式へ変換を行う。これから行う分析、およびデータマイニング・アルゴリズムに必要なデータ形式に合わせて、データの変換を行う。データマイニングで成功するためには、データの完全な分析データ形式をマイニング・アルゴリズムに与えることが重要である。

ステップ 3：データマイニング

ステップ 2.3 で変換したデータを用いてマイニングを行う。これはプロセスの中核をなすステップである。

ステップ 4：結果の分析

ステップ 3 からの出力について解釈および評価を行う。ここで使用する分析方法は、データマイニング操作に応じて様々であり、通常は視覚化手法を含んでいる。

ステップ 5：知識の理解

ステップ 4 で得られた知識の反映。

2.3 地理データにおける知識発見の研究動向

本節では、他の研究者が行っている地理データにおける知識発見の研究、中でも本研究で利用するクラスタリングに絞っていくつか紹介していく。ここで紹介する論文は、それぞれに異なったクラスタリング手法を用いていて、これらは、地理データにおける知識発見の研究の中でクラスタリングを用いた代表的なアプローチである。

まず始めに、本研究にレジムという概念を与えた「Detecting Atmospheric Regimes using Cross-Validated Clustering」というタイトルの論文から紹介する[Smyth 97]。この論文では、北半球のジオポテンシャルな高さの記録の中の低い周期の変わりやすさという大気科学の中で重要なトピックを扱っている。それは、気候モデルと予測に対する深い影響を持っており、その議論は、レジムかどうか、ジオポテンシャル高さの中にクラスタが存在するかどうか、存在するのであれば、そのようなクラスタがいくつあるのかというようなことである。そして、この論文では、どのように cross validation mixture model クラスタリングを行ったかについて述べている。それにより、北半球で 3 つの明確なクラスタが存在するという証拠が得られた。この発見した各クラスタに対する、物理的な説明を行っている。

以上がこの論文の簡単な要約であるが、この論文で用いられている mixture model を使ったクラスタリングについて簡単に付け加えておこう。

d 次元のランダム変数を $\underline{X}$ とし、$\underline{x}$ を $\underline{X}$ の特殊な値を表すと仮定すると、d 構成要素とデータベクトルを観測したもの $\underline{X}$ に対する有限な mixture probability density function は、以下のように表せる。

$$f^{(k)}\left(\underline{x}\middle|\ ^{(k)}\right)=\sum_{j=1}^{k}\ {}_{j}g_{j}\left(\underline{x}\middle|\theta^{(k)}\right)$$

k：モデル中の構成要素の数　　θ_j：g_j（密度構成要素結合）に関するパラメータ

g_j：構成要素密度関数　　α_j：それぞれの構成要素 j に関連した重み

ここで、

$$\sum_j \alpha_j = 1 \text{ and } \alpha_j > 1, 1 \le j \le k$$

である。

そして、

$$\Phi^{(k)} = \{\alpha_1, ..., \alpha_k, \theta_1, ..., \theta_k\}$$

を意味するのが統合的な mixture model に対するパラメータのセットである。

次に「Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications」という論文を紹介する[Sander 98]。この論文では、density-based 概念を用いた GDBSCAN というクラスタリングアルゴリズムについて詳細に説明されている。

density-based クラスタリングのキーアイディアとは、クラスタの各々のポイントに対して、その Eps>0 に対する Eps-neighborhood が、少なくともポイントの最小値以上になる、すなわちポイントの Eps-neighborhood の density がいくつかの閾値を超えなければならないということである。これによって、簡単にポイントのクラスタとこれらのクラスタに属していないノイズポイントを見つけることができる。なぜなら、クラスタの外側より、かなり高いクラスタの中にポイントの典型的な density があり、さらに、ノイズの中にある density は、クラスタのどの density よりも低いからである。

density-based クラスタリングのこのアイディアは、2 つの重要な方法に一般化される。第一に、neighborhood の定義が、相対的で再帰用法である binary に基礎をおくと仮定すると、Eps-neighborhood の代わりに neighborhood の概念を使うことが出来る。第 2 に、オブジェクトの neighborhood でのオブジェクトを単に数える代わりに、他の測定方法を neighborhood の大原則を定義するために使うことが出来る。

最後に、本研究において用いた、"自然な" K 発見のアルゴリズムが記載されている「Efficient and Effective Clustering Methods for Spatial Data Mining」という論文を紹介する[Ng 94]。この論文では、PAM アルゴリズム、CLARA アルゴリズムを改善した CLARANS（Clustering Large Applications based on RANdomized Search）と PAM アルゴリズムや CLARA との計算時間についての比較結果につい

ての説明や CLARANS を空間データマイニングに適用した SD アルゴリズム、CLARANS を非空間データマイニングに適用した NSD アルゴリズムなどが記載されており、それらの効率についても述べられている。

第 3 章
地理データへの
クラスタリング手法の適用

この章では、東アジアにおける酸性･酸化性物質の動態解明に関する有効な知識、特に気候モデルと予測に対し、深い影響力をもっているレジムを発見するという本研究の目的を達成するために、地理データにおける知識発見の先行研究[*1]で多く見られるクラスタリング手法を用いることとした。

クラスタリング手法とは、各々が p 個の変数値をもつ、n 個の対象からなる標本が与えられたとき、”似たもの”が同じクラスになるように、これらの対象をいくつかのクラスに分類する方法を考案する問題を解くためのものである。この方法は完全に数値的でなければならない。また対象が分類されるクラスの数は未知である。

クラスタリングを行う理由として、いくつかのことが考えられる。まず第１に、"真の"グループを見つけ出すことが問題である。第２に、クラスタリングは、データを削減するのに有用である。一方、もしクラスタリングによって予期しない分類が生成された場合には、それ自体が検討されるべき関係であることを示唆している。

本研究では､最も一般的なクラスタリング手法である K-Means クラスタリングと、2.3 で紹介した PAM クラスタリングを用いる。

[*1] http://www.dbs.informatik.uni-muenchen.de/dbs/project/publikationen/veroeffentlichungen.html

本研究において、これらのクラスタリング手法を用いる理由として、この 2 つのクラスタリング手法の共通の特徴である分割数を前もって決められることが挙げられる。これにより、指定した任意の分割数に分類することが可能となる。そして、これは 2.3 でも紹介した"自然な" K 発見アルゴリズム[Ng 94]を用いるためにも必要である。その他にも、PAM クラスタリングで用いられている K-Medid 手法の特徴であるノイズの存在に強く発見されるクラスタが計算の順番によらないということもある。

本章では、K-Means アルゴリズム[ハーティガン 83]と PAM アルゴリズム[Kaufman 90]という２つのクラスタリングアルゴリズムについて詳細に述べ、加えて、"真の"グループを見つけ出す手法[Ng 94]の一つについても述べる。

3.1 K-Means アルゴリズム

準備

j 番目の変数の i 番目のケースが A(i,j)(1≦i≦m, 1≦j≦n)となる値を持つ。変数はケース間のユークリッド距離が適切であるように尺度化される。分割 P(m,k)はクラスタ 1,2, ,k により構成される。m 個のケースのそれぞれは k 個のクラスタ中のどれか１つにある。j 番目の変数の L 番目のクラスタにおけるケース全体に対する平均は、B(L,j)によって表される。クラスタ L におけるケースの数は N(L)である。i 番目のケースと L 番目のクラスタとの間の距離は、

$$D(i,L) = \left(\sum_{j=1}^{n} [A(i,j) - B(L,j)]^2\right)^{1/2}$$

である。分割の誤差は、

$$e[P(m,k)] = \sum_{i=1}^{m} D[i,L(i)]^2$$

である。ここで L(i)は i 番目のケースを含むクラスタである。全般的な手順は、ケースを１つのクラスタから他のクラスタへと移動させることによって、小なる誤差 e をもつ分割を探索するというものである。探索はこのような移動が誤差 e を小さくしなくなった時に終わる。

step1

初期クラスタを 1,2,　,k と仮定する。クラスタ平均 B(L,j) (1≦L≦k, 1≦j≦n)と初期誤差

$$e[P(m,k)]=\sum_{i=1}^{m} D[i,L(i)]^2$$

を計算する。ここで D[i,L(i)]は i と i を含むクラスタのクラスタ平均との間のユークリッド距離を表す。

step2

第１のケースにつき、すべてのクラスタ L に対し次を計算する。

$$\frac{N(L)D(1,L)^2}{N(L)+1}-\frac{N[L(1)]D[1,L(1)]^2}{N[L(1)]-1}$$

これは第 1 のケースの、それが現在属しているクラスタ L(1)からクラスタ L への移動における誤差の増加である。もしすべての L≠L(1)を通じて個の量の最小値が負であれば、第 1 のケースをクラスタ L(1)からこの最小値を持つクラスタ L に移し、L(1)のクラスタ平均と最小値を持つクラスタ L とを調節し、誤差の増加(負)を e[P(m,k)]に加える。

step3

ステップ 2 を i 番目のケース(2≦i≦m)について繰り返す。

step4
もしどのケースについても 1 つのクラスタから他のクラスタへのケースの移動がなければ、停止する。そうでなければ、ステップ 2 に戻る。

3.2 PAM アルゴリズム

PAM(Partitioning Around Medoids)は、K 個のクラスタを見つけるためのアプローチとして、各々のクラスタに対する代表的オブジェクトを決定する。この代表的オブジェクトは Medoid と呼ばれ、クラスタの中で最も中心に位置しているオブジェクトであると意味されている。

PAM アルゴリズムは、BUILD と SWAP と呼ばれるフレーズからなり、最初のフレーズである BUILD は、代表オブジェクトの選択を K 回繰り返すもので、２つ目のフレーズ SWAP は、BUILD によって選択された代表オブジェクトの改善を行うものである。以下に BUILD と SWAP の詳細なアルゴリズムを示す。

BUILD

最初の代表オブジェクトは、すべての他のオブジェクトへの相違の合計が可能な限り小さいものであり、このオブジェクトは、オブジェクト集合の中で最も中心に位置している。残りの(K-1)個の代表オブジェクトの選択方法は次の通りである。

BUILD step1
まだ選択されてないオブジェクト i を考察する。

BUILD step2
選択されてないオブジェクト j を考察し、そして、以前に選択された最も似ているオブジェクトをもっている D_j とオブジェクト i との相違 d(j,i)を計算する。

BUILD step3

もし、この差が正なら、オブジェクト j は、オブジェクト i の選択決定について寄与するだろう。それゆえ、

$$C_{ji} = \max(D_j - d(j,i),0)$$

を計算する。

BUILD step4

オブジェクト i を選択することによって得るトータルゲインを計算する。

すなわち、

$$\sum_j C_{ji}$$

BUILD step5

まだ選択していないオブジェクト i を下の式より選ぶ。

$$\max_i \sum_j C_{ji}$$

このプロセスは、K オブジェクトを見つけるまで続けられる。

SWAP

オブジェクト i が選択されていて、オブジェクト h が選択されていない。オブジェクト(i,h)のすべてのペアを考慮する。

SWAP step1

選択してないオブジェクト j を考察し、そして、SWAP まで寄与 C_{jih} を計算する。

SWAP step1.1

もし、j が他の代表オブジェクトの一つからよりも i と h の両方から遠いなら、C_{jih} は 0 である。

SWAP step1.2

もし、j がいくつかの選ばれた代表オブジェクト($d(j,i)=D_j$)からより i から遠くにないなら、2 つの状況を考えなければならない。

SWAP step1.2a

j は 2 番目に近い代表オブジェクトより h に近い。

$$d(j,h) < E_j$$

ここで E_j は j と 2 番目に最も近い代表オブジェクトの間の相違である。このケースで、オブジェクト i と h の間の SWAP へのオブジェクト j の寄与は、

$$C_{jih} = d(j,h) - d(j,i)$$

である。

SWAP step1.2b

j は少なくとも 2 番目に近い代表オブジェクトより h から遠い。

$$d(j,h) \geq E_j$$

このケースで、SWAP へのオブジェクト j の寄与は、

$$C_{jih} = E_j - D_j$$

である。

ステップ 1.2a において寄与 C_{jih} はオブジェクト j,h,i の関係している位置によって正にも負にもなり得る。

SWAP step1.3

j が少なくとも他の代表的オブジェクトの一つよりもオブジェクト i から遠く、いくつかの代表オブジェクトより、h に近い場合、SWAP への j の寄与は、

$$C_{jih} = d(j,h) - D_j$$

である。

SWAP step2

C_{jih} を加算することによって、SWAP のトータル結果を計算する。

$$T_{ih} = \sum_j C_{jih}$$

次のステップで、SWAP を実行するかどうか決める。

SWAP step3

下記の式を満たすペア(i,h)を選択する。

$$\min_{i,h} T_{ih}$$

SWAP step4

もし、最小の T_{ih} が負なら、SWAP を実行しステップ 1 へ戻る。もし、最小の T_{ih} が正か 0 なら、オブジェクトの値は SWAP の実行によって減少出来ない、よって、アルゴリズムを停止する。

3.3 "自然な" K の発見的手法

この節では、"自然な" K の発見的手法について説明する。まず、"自然な" K の発見的手法に必要なシルエット幅・シルエット係数の説明を行う。

シルエット幅 s(i)の計算方法を以下に示す。

初めに s(i)を定義するための方法として、オブジェクト i がクラスタ A に属していると仮定する。a(i)をオブジェクト i とクラスタ A の他のオブジェクトまでの平均相違とする。ここで注意すべきことは、クラスタ A に属しているオブジェクトは、１つではないということである。

そして、クラスタ A から異なったクラスタ C を考え定義する。ここで、d(i,C)をオブジェクト i からクラスタ C の全てのオブジェクトへの平均相違とする。クラスタ C がクラスタ A と異なっているとき、d(i,C)を計算し、その最小値を選択する。

つまり、

$$b(i) = \min_{C \neq A} d(i,C)$$

を計算する。この最小を達成されたクラスタ B は、オブジェクト i の近くにあるということになる。これは、オブジェクト i に対する二番目にベストな選択である。もし、クラスタ A が捨てられたなら、クラスタ B が i へ最も近い。

これらから s(i)は、以下の式で計算される。

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

これによって s(i)は、$-1 \leq s(i) \leq 1$ の範囲の値をとることとなる。ここで s(i)の値が負であったらなば、その s(i)の値を正の値に変換し、s(i)の範囲を $0 \leq s(i) \leq 1$ とする。

そして、オブジェクト i が属するクラスタにおいて、s(i)の平均をとったものが、そのクラスタの平均シルエット幅であり、すべてのオブジェクトについて平均をとったものが、そのデータ集合の平均シルエット幅である。

K の値を変えて計算を行った結果、以下の式を満たすものが、シルエット係数と呼ばれるものである。

$$SC = \max_{K} \bar{s}(K)$$

SC	説明
0.71-1.00	強い構造のクラスタ
0.51-0.70	ほどよい構造のクラスタ
0.26-0.50	弱い構造のクラスタ
≦ 0.25	十分な構造を持ったクラスタがない

表 3.1 シルエット係数とクラスタの関係

次に、本研究でも用いた"自然な" K を発見するアルゴリズムを紹介する。

"自然な" K の発見的手法

"自然な" K の発見的手法 step1
最も高いシルエット係数により K の値を見つける。

"自然な" K の発見的手法 step2
もし、全ての K クラスタが、シルエット幅≧0.51 を持っていたら、BestK=K とし、終了。

"自然な" K の発見的手法 step3
一方、シルエット幅が 0.5 以下のクラスタに属するオブジェクトを取り除く。取り除かれたオブジェクトは、ノイズであるとみなされる。そして、ノイズなしの新しいデータセットのためにステップ 1 に戻る。

"自然な" K の発見的手法 step4
もし、ステップ 3 で取り除かれたノイズの数が閾値（例えば,オブジェクトのトータル数の 25%）を超えるなら、単純にどのクラスタリングも適さないという結果を示している。ここで、BestK=1 とし終了。

第 4 章
設計 & 実装

本章では、各アルゴリズムを実装するまでの要求仕様・機能仕様・コード設計について順に述べていく。その前にそこで用いている語句の簡単な説明を行う。

要求仕様とは、必要とするプログラム動作をごく一般的な言葉を使って記述するものであり、機能仕様とは、プログラムの機能の詳細を記述したものである。そして、コード設計は、主要アルゴリズム、モジュール仕様、データ構造についての設計を行うものである。

本研究では、UNIX 上で C 言語を用いて、実装を行った。表 4.1 にコメントを含む各プログラムの大まかなコード数を示す。

クラスタリング手法	コード数
K-Means	600 行程度
PAM & "Best" K	1200 行程度

表 4.1 各プログラムのコード数

4.1 K-Means クラスタリング

4.1.1 要求仕様

求めたい分割数を与えることによって、データの分割を行う。"自然な" K を見つけるための判断基準となる統計的な結果を出力する。

4.1.2 機能仕様

以下にプログラムの機能を列挙する。

・データ入力
・与えられた分割数により、データを分割
・どのような条件（オブジェクト数、分割数、誤差）で計算を行ったかを出力
・それぞれのクラスタ間の距離を出力
・それぞれのクラスタの平均値、標準偏差、分散を出力
・クラスタごとに、属しているオブジェクトの出力

4.1.3 コード設計

主要アルゴリズム

3.1 参照

モジュール仕様

主要なモジュールに kmeans( )というものがあるが、このモジュールでは、K 個の初期クラスタを獲得し、誤差の計算により改善を行っている。他に入力モジュールと簡単な統計処理も行う出力モジュールがある。

データ構造

2 次元配列に格納

4.2 PAM クラスタリング&"自然な" K の発見

4.2.1 要求仕様

求めたい分割数を範囲として与えることによって、データを分割する。そして、その範囲の中で"自然な"分割数を求める。

4.2.2 機能仕様

以下にプログラムの機能を列挙する。

・データ入力
・与えられた範囲で分割
・"自然な"分割を求める
・"自然な"分割が見つかった場合に、いくつかのデータを出力
・計算時間の出力

4.2.3 コード設計

主要アルゴリズム

PAM アルゴリズムに関しては、4.2 を参照。"自然な" K の発見に関しては、4.3 を参照。

モジュール仕様

主要なモジュールに PAM( )、Build( )、Swap( )、BelongTo( )、FindBestK( )、runtime( )がある。以下にそれぞれのモジュールについての簡単な説明を行う。

PAM( )：
このプログラムの核であり、このモジュールの中で、Build( )、Swap( )、BelongTo( )、FindBestK( )が含まれている。

Build( )：
PAM アルゴリズムの BUILD に対応するモジュール。

Swap( )：
PAM アルゴリズムの SWAP に対応するモジュール。

BelongTo( )：
オブジェクトがどのクラスタに属しているかを計算するモジュール。

FindBestK( )：
"自然な" K の発見的手法に対応するモジュールである。

runtime( )：
プログラムの実行時間を測定するモジュールである。

データ構造

```
struct data{
   int lengthl;
   int lengthr;
   double **DatAry;
};
```

int lengthl：　　データの行数
int lengthr：　　データの列数
double ＊＊DatAry：　動的な 2 次元配列に格納

```
struct datagroup{
   int *obj;
   int *belong;
   double *S;
   double *EachSW;
   double SC;
};
```

int *obj;	代表オブジェクトを格納する1次元配列
int *belong;	オブジェクトがどのクラスタに属しているかを示す1次元配列
double *S;	各オブジェクトのシルエット幅
double *EachSW;	各クラスタのシルエット幅(平均)
double SC;	その分割数でのシルエット係数

第 5 章
実験

5.1 データの準備

データの準備は、プロセスの中で最も資源を消費するステップであり、一般にプロジェクト全体の労力の最高 60%を必要とする。このステップは、次の 3 つの段階から構成されている。

1. データの選択（データの識別と抽出）
2. データの前処理（データ･サンプリングと品質チェック）
3. データの変換（分析用モデルへのデータ変換）

5.1.1 データの選択

データの選択の目的は、使用可能なデータ･ソースを識別すること、またマイニングをさらに続けるための準備として、予備的な分析を行うのに必要なデータを抽出することである。

本研究では、2.2 でも述べたが、地球環境研究センターによって収集･整理された「'97IGAC/APARE/PEACAMPOT 航空機･地上観測データ('91～'95 集成版)」の"フライトデータ"とそれに対応する"IGAC/APARE/PEACAMPOT 航空機調査におけるオゾン（O_3）、窒素酸化物(NO_x)、および二酸化硫黄（SO_2）の観測"を選択した。この

データは数値データとしてまとめられている。

5.1.2 データの前処理

データの前処理の目的は、選択したデータの品質を確保することである。クリーンでよく整備されているデータは、他の定量分析の場合と同様に、データマイニングを成功させるための明確な前提条件である。さらに、手元にあるデータに一層精通することにより、このマイニング段階で、実際の知識をどこから見つけるべきか、よくわかるようになる。

本研究では、データの前処理として、欠損値を取り扱った。欠損値とは、選択したデータに含まれていない値や、ノイズの検出の際に削除した可能性のある無効値などのことである。値の欠損は、人為的ミスのため、入力時に情報が入手できなかったため、あるいは複数の異質ソースからデータが選択されて、ミスマッチが生じたために起こることがある。欠損値を取り扱うため、データ分析者は種々の手法を用いるが、どの手法も完全なものではない。次にいくつかの欠損値の取り扱いについて簡単に述べる。

まず、欠損値を含んでいる観測結果を除外するだけという手法がある。これは簡単だが、明らかに貴重なデータを失うという欠点がある。このデータ消失は、データ量が多ければさほど問題ではないかもしれないが、少量データのマイニングの結果や詐欺や品質管理が目標である場合の結果には確実に影響を与える。このような状況では、見つけようとしていた、まさにその観測結果を捨ててしまう可能性が十分にある。実際に、値が欠損しているという事実は、詐欺または品質の問題の原因を探る手がかりとなる可能性が十分にあるからである。

かなりの数の観測結果において、同じ変数に欠損値がある場合には、分析から当該変数を除外することができるかもしれない。しかし、この場合も重大な結果が生じる。分析者に知られなかった当該変数が、ソリューションの重要な鍵となり得るからである。

従って、データの観測結果または変数（あるいはその両方）を除外するという決定は、安易に行うべきではないし、その結果を予測することも容易なことではない。幸いにも、欠損値の問題に関しては、いくつかの方法がある。その 1 つが、欠損値

を最も可能性の高い値に置きかえるという方法である。数値変数の場合、このもっとも可能性の高い値として、平均値または最頻値を使用することができる。カテゴリー変数の場合には、最頻値を使用することもできるし、あるいは当該変数に関して新しく作った値、たとえば UNKNOWN という値を使用することもできる。

また、数値変数とカテゴリー変数の両方について使用できる最も高度な方法がある。この方法では、観測結果の他の変数の値をもとに、変数が取りうる可能性が高い値を予測モデルを用いて予測する。

このように欠損値問題の対処方法はいくつかあるが、平均値や予測を用いる方法は、すべて犠牲を伴うということを心得ておかなければならない。推測が必要であればあるほど、データベースは実際のデータとはかけ離れていく。そのことは、データマイニング結果の精度や妥当性にすぐさま影響を及ぼす。

本研究で行った欠損値の取り扱いは、EXCEL を用いて、欠損値を含んでいる観測結果の除外を行った。その後、それらのデータを PostgreSQL（フリーの RDBMS）上に Perl を用いた簡単なスクリプトを作成し、登録を行った。

5.1.3 データの変換

5.1.2 によって、PostgreSQL 上に登録された 10 個のデータベースを SQL 言語を用いて、1 つのデータベース（Earth）へと結合した。この Earth データベースの大きさは、2000 ケース×12 カラムである。

表 5.1 に Earth データベースのサンプルを示す。

年	月	日	北緯	東経	高度
1992	11	8	32.52	129.43	5575.5
1992	11	8	32.51	129.4	6493.5
1992	11	8	32.51	129.37	7548
1992	11	8	32.5	129.33	8397
1992	11	8	32.49	129.3	9168

速度	温度	湿度	オゾン	二酸化硫黄	窒素
303.1724	9	47.6	40.9	0.92	0.02
317.8032	7.1	39.5	42.2	1.25	0.35
310.3952	5.6	26.3	43.5	1.27	0.33
306.1356	6.9	6.3	44.6	0.71	0.03
300.5796	7.6	0.8	44.6	0.27	0.17

表 5.1 Earth

5.2 マイニング

このステップの目的は、選択したデータマイニング･アルゴリズムを前処理されたデータに的確に適用することである。

本研究では、5.1 で作成した Earth データベースに対しクラスタリングを行う。マイニングのステップでは、PAM クラスタリング& "自然な" K の発見的手法を用いてクラスタリングを行う。

表 5.4 は、マイニングを行った際の実行環境とその主な結果である"自然な" K についてまとめたものである。

使用計算機	Load Sharering Facility: cs1		
データベース名	K の範囲	計算時間	Best K
Earth	2-50	29h 44min 21sec	39

表 5.2 実行環境と"自然な" K

5.3 結果の分析

言うまでもなく、マイニング処理の結果を分析することは、プロセス全体における最も重要なステップの 1 つである。また、マイニング結果をグラフィック表示することも分析をする上で重要である。

表 5.3 に Earth データベースに対してマイニングを行った結果である、各クラスタのシルエット幅と属しているオブジェクトの数をまとめたものである。

表 5.3 のマイニングの結果は、あまりにも"自然な" K の値が大きすぎるので、2 つのステップからなる、以下の手法を用いて分析するデータを減らす。

step1

ノイズを除いたデータをマイニング結果である"自然な"K の値で割り、小数点以下を切り捨てる。その値を仮に判定基準オブジェクト数を呼ぶことにする。そこで、判定基準オブジェクト数を含み、かつその数より少ないオブジェクトを持つクラスタを取り除く。

step2

"自然な"K の際の平均シルエット幅より、小さいシルエット幅を持つクラスタを取り除く。

表 5.4 に、上記のアルゴリズムを用いて計算した、分析するデータを減らす際の判定基準を示し、表 5.5 にノイズとして除去されなかった残りのクラスタを示す。

Cluster Number	Silhouette Width(Ave.)	Object Number
1	0.98953	35
2	0.964244	77
3	0.995368	94
4	0.988143	143
5	0.924479	58
6	0.97267	101
7	0.984855	86
8	0.965309	36
9	0.888368	89
10	0.916153	13
11	0.912104	63
12	0.953401	91
13	0.740408	25
14	0.973758	52
15	0.933688	49
16	0.945283	39
17	0.963323	38
18	0.909649	28
19	0.886969	33
20	0.94605	18
21	0.814346	8
22	0.916164	14
23	0.924256	31
24	0.856304	21
25	0.957636	150
26	0.902221	48
27	0.658546	17
28	0.96603	22
29	0.925459	30
30	0.974189	56
31	0.705853	11
32	0.933124	54
33	0.59003	7
34	0.869646	19
35	0.968363	94
36	0.962021	58
37	0.724307	5
38	0.997229	107
39	0.895545	9

表 5.3 各クラスタのシルエット幅と属しているオブジェクトの数

	オブジェクト数	平均シルエット幅	"自然な" K	オブジェクト数 / "自然な" K
使用	1929	0.905000538	39	49
ノイズ	19			
合計	1948			

表 5.4 判定基準

Cluster Number	Silhouette Width(Ave.)	Object Number
2	0.964244	77
3	0.995368	94
4	0.988143	143
5	0.924479	58
6	0.97267	101
7	0.984855	86
11	0.912104	63
12	0.953401	91
14	0.973758	52
25	0.957636	150
30	0.974189	56
32	0.933124	54
35	0.968363	94
36	0.962021	58
38	0.997229	107

表 5.5 データを減らした後の各クラスタの
シルエット幅と属しているオブジェクトの数

表 5.5 のクラスタを地理的（2 次元：北緯、東経）にプロットすると図 5.1 のようなものになる。

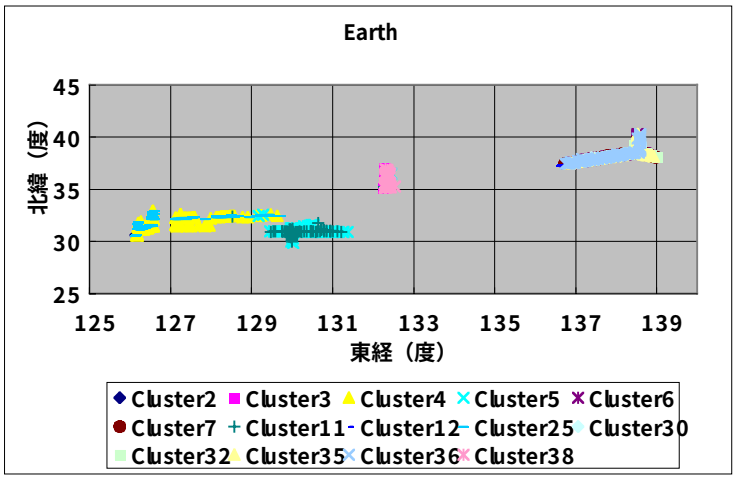


図 5.1 各クラスタの 2 次元上の分布 1

ここで、図 5.1 を見てみると図 5.2 に示すように分割することができる。

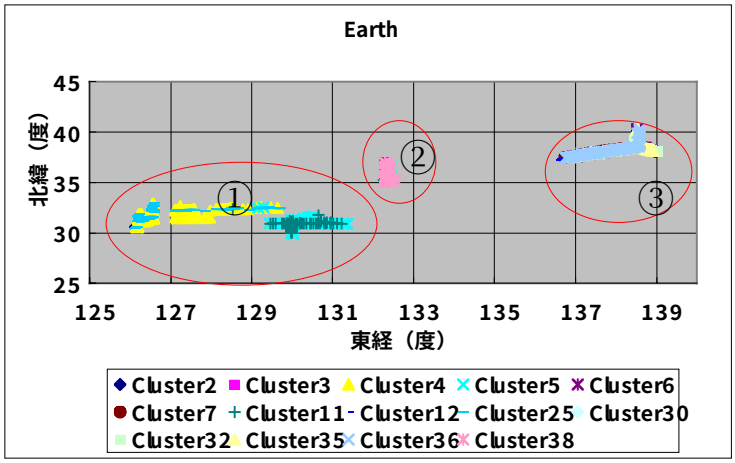
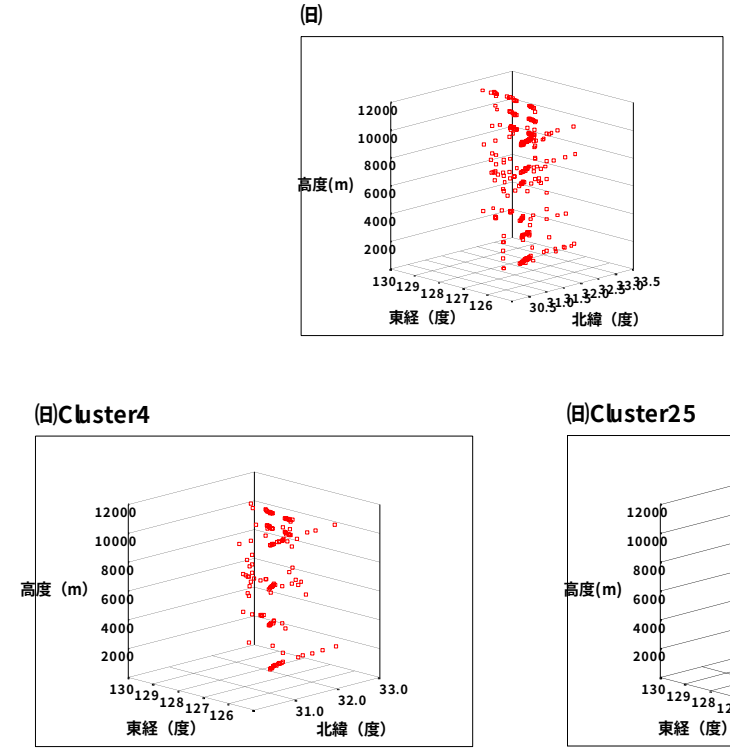


図 5.2 各クラスタの 2 次元上の分布 2

図 5.2 を見る上で重要となってくるが高さという次元についてである。分割①・②・③における 3 次元プロット（北緯、東経、高度）を図 5.3、図 5.4、図 5.5 に示す。



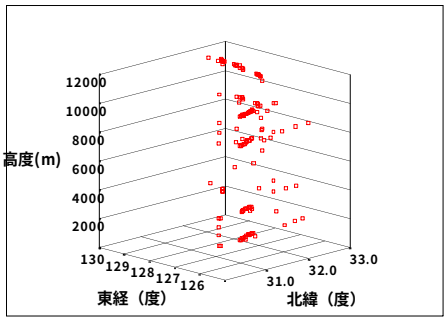
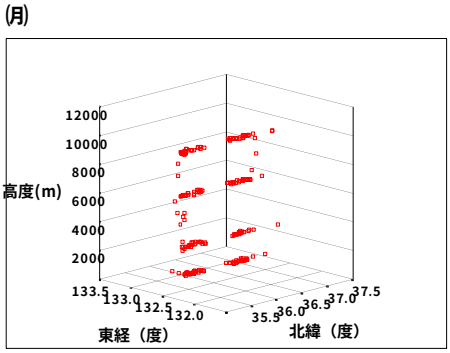
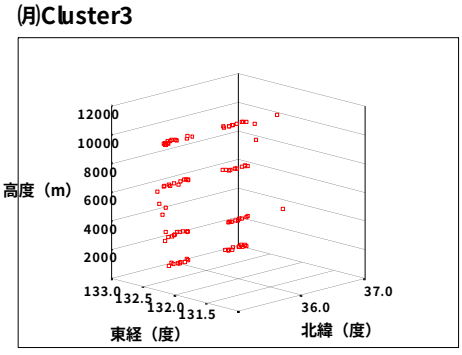


図 5.3 分割①における 3 次元プロット





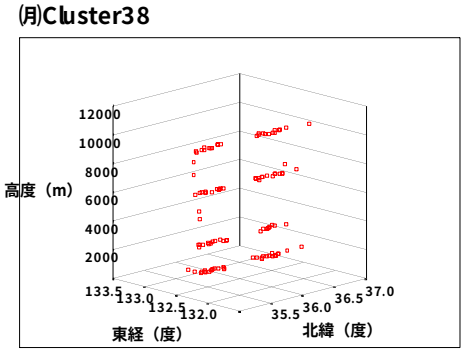
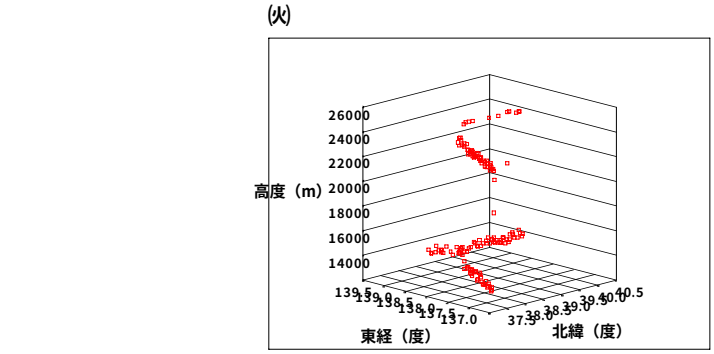


図 5.4 分割②における 3 次元プロット



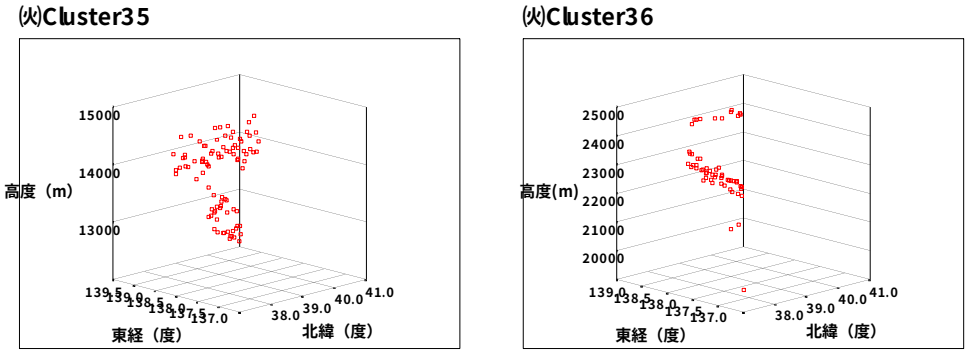


図 5.5 分割③における 3 次元プロット

ここで、③は高度によってきれいに 2 つに分割されていることがわかる。よって、③を③と④に分割する。

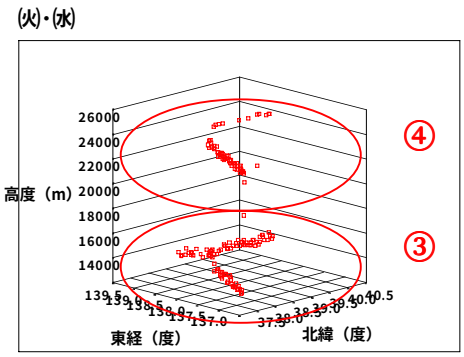


図 6.6 分割③･④

これらの視覚的分析により、Earth データベースにおける"自然な"分割は、4 つであると考えられる。下表 5.6 は分割数 4 での代表クラスタにおける各属性値を表している。

	クラスタ番号	温度（度）	湿度（度）	オゾン(ppb)	二酸化硫黄(ppb)	窒素(ppb)
①	Cluster4	9.2227	33.3421	44.9116	0.3132	0.7422
	Cluster11	2.1284	31.7226	44.7244	0.9188	1.1921
②	Cluster3	8.5666	28.3110	44.3662	1.1748	1.3579
	Cluster38	8.1011	28.7453	44.6878	0.9646	0.9786
③	Cluster35	3.2975	54.1828	35.2633	0.3843	0.6318
④	Cluster36	1.5277	58.0002	34.2926	0.3566	0.5755

表 5.6 分割数 4 での代表クラスタにおける各属性値

ここで、表 5.6 から次のような特徴を得ることができる。

1.各分割を比較すると、分割②が最も汚染されている。
2.同じような立体構造を持つ分割①において、代表クラスタ間の温度と二酸化硫黄の関係が異なっている。

表 5.7 に分割①における温度と二酸化硫黄の相関分析を行った結果を記載する。

記述統計量

	平均値	標準偏差	N
温度	9.2227	3.8522	126
硫黄	.3132	.2976	126

記述統計量

	平均値	標準偏差	N
温度	2.1284	5.3037	135
硫黄	.9188	.3809	135

相関係数

		温度	硫黄
Pearson の相関係数	温度		
	硫黄	.111	
有意確率 (両側)	温度		
	硫黄	.216	
平方和と積和	温度		
	硫黄	15.898	
共分散	温度		
	硫黄	.127	
N	温度		
	硫黄	126	

相関係数

		温度	硫黄
Pearson の相関係数	温度		
	硫黄	.431**	
有意確率 (両側)	温度		
	硫黄	.000	
平方和と積和	温度		
	硫黄	116.645	
共分散	温度		
	硫黄	.870	
N	温度		
	硫黄	135	

**. 相関係数は1%水準で有意 (両側) です。

表 5.7 分割①における温度と二酸化硫黄に関する相関分析結果

クラスタ 4(左)とクラスタ 11(右)

相関分析とは、相関係数を測って、相関の程度を判断することである。ここで相関係数とその有意性について簡単に説明を行う。

相関係数とは、2 つの変量が x と y のとき、相関係数 r は

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_N - \bar{x})(y_N - \bar{y})}{\sqrt{\{(x_1 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2\}\{(y_1 - \bar{y})^2 + \cdots + (y_N - \bar{y})^2\}}}$$

と定義される-1≦r≦1 の範囲の値である。

次に、相関係数の有意性について説明を行う。

相関係数は分布を行い、自由度によって変わる。相関係数の分布の自由度(d.f.: degree of freedom)は、d.f. = 標本サイズ - 変数の数である。相関係数の分布の上で、信頼係数 95%と 99%のケースについて、相関係数の有意水準を、自由度ごとに表にしたものが表 6.8 である。

自由度	5%	1%
1	0.997	1.000
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.874
100	0.195	0.254
125	0.174	0.228
150	0.159	0.208

表 5.8 相関係数の 5%の有意水準および 1%の有意水準[鳥居 94]

計算した相関係数が表 5.8 の相関係数より高ければ、有意な相関があるという。

これらから、クラスタ 11 において、温度と二酸化硫黄の間には有意な相関関係が

あることがわかる。

5.4 考察

この節では、マイニング結果の分析によって得た次の 3 つの事象について考察を行う。

1. Earth データベースにおける"自然な"分割数は、4 つであること。
2. 各分割を比較すると、分割②が最も汚染されていること。
3. 同じような立体構造を持つ分割①において、代表クラスタ間の温度と二酸化硫黄の関係が異なっている。そして、クラスタ 11 において温度と二酸化硫黄の間には有意な相関関係があること。

まず、1.について考察を行う。

5.2 のマイニングで行ったクラスタリングにより得た"自然な"分割数は、39 であった。しかし、データの地理情報に関しての散布図を用いて、クラスタリング結果を視覚的にすると、分割数 4 が最も"自然"であると考えることができる。それでは、分割数 4 の際のクラスタは、レジムなのであろうか。ここでレジムとは、大気データセットから見分けることができる、循環し永続的な空間パターンのことである。分割数 4 の際のクラスタは、空間パターンではあるが、このデータからは、循環･永続的とは言えない。よって、分割数 4 のクラスタはレジムであるとは言えない。

以下に、本研究で用いた"自然な" K 発見アルゴリズムの問題点をあげる。

・"自然な"分割数がオブジェクト数に比例して、大きくなる傾向がある。これはいわば当然といえば当然のことかもしれない。本研究において"自然な"分割数は 39 となったが、これでは分割数が大きすぎてデータの分析･理解が難しい。

・アルゴリズム的に計算量が膨大である。これは、ノイズデータを検出した際にまた新たにデータセットを作り直し、最初から計算をし直すというアルゴリズムであるために計算量が膨大になると考えられる。本研究での計算時間は約 30 時間となったが、これはデータベースの大きさから考えると非常に非効率であると思われる。

以上の点をふまえて、"自然な" K 発見アルゴリズムを作成する必要があると思われる。

次に、2.について考察を行う。ここでは、国立環境研究所により提供された結果を用いて、本研究で得た結果の考察を行う。まず、図 5.7 を見てみると、これは国立環境研究所によって提供された、酸性雨原因物質の輸送パターンを示したものである[Hatayama 97]。この輸送パターンを本研究のデータと照合した結果を図 5.8 に示す。

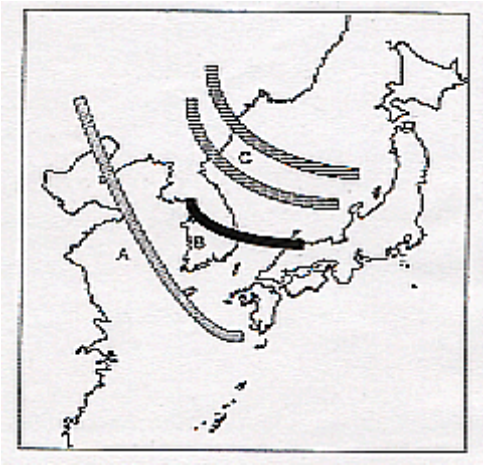


図 5.7 酸性雨原因物質の輸送パターン

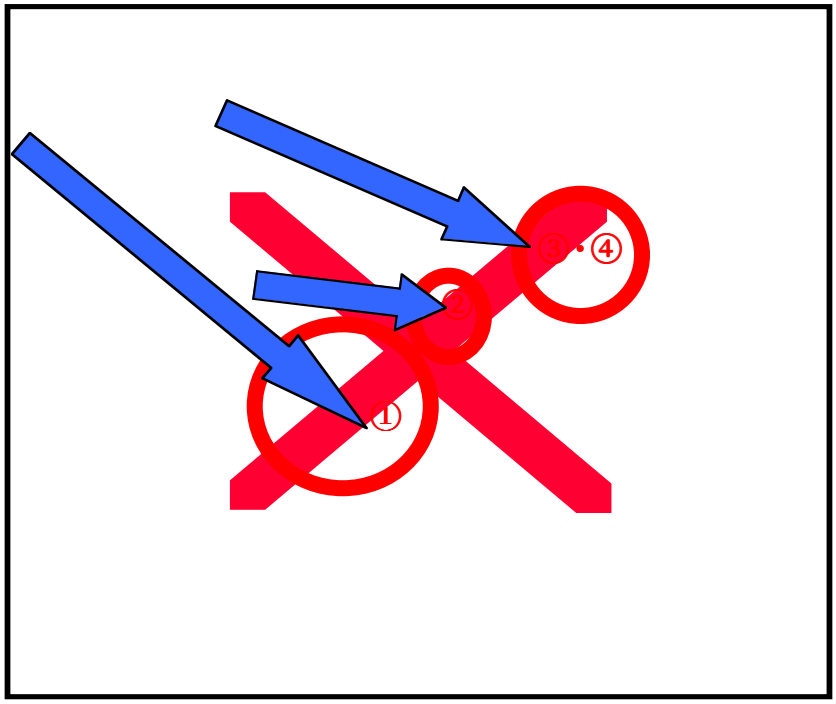


図 5.8 酸性雨原因物質の輸送パターンと本研究データの照合

図 5.8 を考察する上で必要となってくるデータに、アジア地域の酸性雨原因性物質の発生量というものがある。これは、Foell と Green によって、日本などを除く、インドを含めたアジアの国々の硫黄酸化物と窒素酸化物の発生量を 1986 年、2000 年、2010 年の 3 つの時期に対して予測したデータがある。この地域全体の、1986 年の硫黄酸化物の放出量は、二酸化硫黄換算で約 2800 万 t となり、中国が 67%を占め、インドの 11%がこれに続いている。別の見積もりで日本を比較すると、1990 年の排出量は中国 2095 万 t、日本 98.9 万 t、韓国 161.1 万 t である。窒素酸化物についてはこの地域全体で 1986 年には窒素酸化物換算で 1400 万 t の放出がある。中国は 54%を占め、やはりインドの 20%がこれに次いでいる。別の見積もりでは、1990 年の排出量は、中国 672 万 t、日本 160 万 t、韓国 93 万 t である[安成 99]。

このデータから、日本も含め中国･韓国は、酸性雨原因性物質の発生量が多いことがわかる。この大量の酸性雨原因性物質が図 5.7･図 5.8 に示すような経路で日本に向かって移流していることがわかる。ここで問題となるのが、分割②の汚染が他の分割に比べてひどいので、中国から流れてくる分割①･③･④の方が酸性雨原因性物質の発生量が多いのだから汚染がひどくなるのではないかと考えられる。これについてのデータが表 5.9 である[河村 88]。

物質	滞留時間（対流圏）
オゾン	1 ヶ月から 4 ヶ月
二酸化硫黄	数日
酸化窒素	数日から 1 ヶ月以下

表 5.9 酸性雨原因性物質の滞留時間

これらから、2.についての考察をまとめると、中国･韓国などから大気中に放出された酸性雨原因性物質は、風によって風下に運ばれながら、周囲の清浄な空気と混ざって広がり希釈される。そのとき、物質によっては物理的･科学的に変化し、一部は重力によって降下し、また地表面に付着したり降水に洗われたり大気中から除去される。この過程が韓国の方が短いのである。よって、韓国からの酸性雨原因性物質の輸送経路にある分割②の汚染の方が、他の分割よりもひどいのである。

最後に、3.について考察する。これは、同じような立体構造を持つ分割①において、代表クラスタ間の温度と二酸化硫黄の関係が明らかに異なっている。そして、クラスタ 11 においては、温度と二酸化硫黄の間には有意な相関関係がある。

これから、二酸化硫黄は、地球寒冷化の重要な因子である硫酸塩の前駆体として重要な役割を果たしているので、この結果は、地球の寒冷化のシステムの鍵を握る証拠となる可能性があることが考えられる。

第 6 章
おわりに

本論文では、5 章に渡って、地理データへのクラスタリング手法の適用について、用いたクラスタリング･アルゴリズムの説明、実装、実験、考察までを議論してきた。

第 1 章では、本研究の背景･目的について述べた。第 2 章では、地球環境問題の現状について簡単に述べ、本研究で扱うデータについての説明を行った。そして、第 6 章で用いた一般的な KDD のプロセス過程について述べ、地理データにおける先行研究を紹介した。第 3 章では、本研究で用いた地理データへのアプローチである、クラスタリング手法について触れ、K-Means アルゴリズム、PAM アルゴリズム、そして、クラスタリングを行う際に重要となってくる"自然な" 分割数を求めるアルゴリズムについて詳細に述べた。第 4 章では、クラスタリング手法の設計･実装する際の要求仕様、機能仕様、コード設計について説明を行った。第 5 章では、KDD のプロセス過程に沿って分析･考察を行い、次の結果を得た。

1. Earth データベースに対しての"自然な" K の値は、4 であること。
2. 酸性雨原因性物質の輸送距離によっても濃度の違いがあること。
3. 1 つの分割で観測された温度と二酸化硫黄の関係には、有意な相関関係があることにより、地球の寒冷化システムの鍵を握る証拠となる可能性があること。

上記の 2･3 は、本研究によって新しく発見された関係であり、特に 3 においては、今後の大気科学の分野に貢献できるのではないかと思われる。

今後の問題点を挙げると、

・地球環境問題についてのより詳細な調査を行うこと。

・レジムを発見するために、時間の前後関係を考慮した"動的"クラスタリング･アルゴリズムの開発。

・欠損値の識別やノイズデータの除去への視覚的アプローチ、特に 3 次元を超えるデータの場合。

・より"自然な" K を見つけるアルゴリズムの開発。

・大規模データベースに実時間内で用いることができるように、データのサンプリング、並列処理を用いたクラスタリング･アルゴリズムの開発。

が挙げられる。

謝辞

本研究を進めるにあたり、ホーツーバオ教授、石崎雅人助教授,グエンゴクビン助手からは、終始変わらぬご指導を頂きました。心から感謝いたします。

中森義輝教授からは、多大なる助言を頂きました。深く感謝いたします。

知識創造論講座の方々からは、日頃から様々な刺激をいただきました。深く感謝いたします。

参 考 文 献

[Hatayama 95a] Hatayama, S., Murano, K., Bandow, H., Mukai, H. and Akimoto, H.: High Concentration of SO2 Observed over the Sea of Japan.: TAO, Vol.6, No.3, pp.403-408(1995).

[Hatayama 95b] Hatayama, S., Murano, K., Bandow, H., Sakamaki, F., Yamato, M. and Akimoto, H.: The 1991 PEACAMPOT aircraft observation of ozone, NO_x, and SO_2 over the East China Sea, the Yellow Sea, and the Sea of Japan, Jornal of Geophysical Reseach,.Vol.100,. No.D11, pp.23,143-23(1995).

[Hatayama 97] Hatayama, S., Murano, K., Mukai, H., Sakamaki, F., Bandow, H., Watanabe, I., Yamato, M., Tanaka, S. and Akimoto, H.: SO_2 and Sulfate Aerosols over the Seas between Japan and the Asian Continent, エアロゾル研究, 第 12 巻, 第 2 号, pp.91-95(1997).

[Smyth 97] Smyth, P., Roden, J., Ghil, M. and Ide, K.: Detecting Atmospheric Regimes using Cross-Validated Clustering, KDD-97, pp.61-66(1997).

[Sander 98] Sander, J.,Easter, M., Kriegel, H. P. and Xu, X.: Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Application, Data Mining and Knowledge Discovery 2, pp.169-194(1998).

[Ng 94] Ng, R. T. and Han, J.: Efficient and Effective Clustering Methods for Spatial Data Mining, Proceedings of the Twentieth International Conference on Very Large Databases, pp.144-155(1994).

[エイドリアン 98] エイドリアン,ザンティンジ（山本英子,梅村恭司 訳）：データマイニング,共立出版株式会社,(1998).

[ビーガス 97] ビーガス（株式会社 社会調査部他 訳）：ニューラルネットワークによるデータマイニング,日経 BP 社,(1997).

[キャベナ 99] キャベナ,ハジリアン,スタッドラー,ベルフィーズ,ザナシー（日本アイ･ビー･エム株式会社・ナショナル･ランゲージ･サポート 訳）：データマイニング活用ガイド,株式会社トッパン,(1999).

[安成 99] 中澤高清,原宏,住明正,森田恒幸,米本昌平(安成哲三,岩坂泰信 編):大気環境の変化,岩波書店,(1999).

[河村 88] 河村武,新藤静夫,田瀬則雄,吉田富雄,高野健三,手塚敬祐,石塚皓造,藤原喜久夫, 岩城英夫,藤井宏一,高橋正征（河村武,岩城英夫 編）:環境科学Ⅰ 自然環境系,朝倉書店,(1988).

[ハーティガン 83] ハーティガン（西田春彦,吉田光雄,平松闊,田中邦夫 訳）：クラスタ分析,マイクロソフトウェア株式会社,(1983).

[Kaufman 90] Kaufman, L. and Rousseeuw, P. J.: Finding Groups in Data, John Wiley & Sons.,(1990).

[鳥居 94] 鳥居泰彦：はじめての統計学,日本経済新聞社,(1994).

[Rodhe 89] Rodhe, H.: Acidification in a global perspective., Ambio, 18, pp.155-160(1989).

[Galloway 89] Galloway, J. N.: Atmospheric acidification: projections for the future., Ambio, 18, pp.161-166(1989).

[地球環境センター 96] 地球環境センター：'95 IGAC/APARE/PEACAMPOT 航空機･地上観測データ集,環境庁 国立環境研究所 地球環境研究センター(1996).

付録