階層的クラスタリングを用いた台風被害予測モデルの構築手法

中村 翔† 森 康真‡ 北上 始‡

†広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1
‡広島市立大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1
E-mail: †sho@de.info.hiroshima-cu.ac.jp, ‡{mori, kitakami}@hiroshima-cu.ac.jp

あらまし 本論文では，台風の被害予測を行うために，デジタル台風の Web サイトで提供されている総合データベースの中から台風データベースと被害データベースを用いて，台風被害予測モデルを構築する方法を提案する．台風被害予測モデルを構築するために，台風データベースと被害データベースのそれぞれに含まれるデータを分類し，両者の分類結果を統合している．台風データベースに含まれるデータの分類については，台風の経路や移動速度という観点で実施されおり，被害データベースに含まれるデータの分類については，pLSI による次元縮約を考慮した階層的併合クラスタリングにより実施されている．また，提案手法に対する被害予測精度を評価したので，その評価結果についても報告する．

キーワード データマイニング，階層的クラスタリング，次元縮約，デジタル台風

A Construction Method of a Typhoon Damage Prediction Model Using Hierarchical Clustering Method

Sho NAKAMURA† Yasuma MORI‡ and Hajime KITAKAMI‡

†Faculty of Information Sciences, Hiroshima City University 3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima-Shi, Hiroshima, 731-3194 Japan
‡Graduate School of Information Sciences, Hiroshima City University 3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima-Shi, Hiroshima, 731-3194 Japan
E-mail: †sho@de.info.hiroshima-cu.ac.jp, ‡{mori, kitakami}@hiroshima-cu.ac.jp

Abstract In order to predict typhoon damage for each district in prefecture, this paper proposes a method for constructing the prediction model of typhoon damage using both route and damage databases related to typhoon, where the both are supplied from the website of Digital Typhoon. The construction of the prediction model is achieved by integrating between both typhoon route pattern and damage pattern sets, where the both are extracted from typhoon databases and damage databases, respectively. The typhoon route pattern set is represented as the combination of both a typhoon route and speed. The damage pattern set is extracted by agglomerative hierarchical clustering using the dimension reduction based on pLSI. Finally, we report experimental results for the evaluation of damage prediction precision for the typhoon damage model.

Keyword datamining, hierarchical clustering, dimension reduction, digital typhoon

1. はじめに

地球科学，生物情報学，物理学などの分野に登場する巨大科学データベースの分野では，継続的に生み出される爆発的な量のデータから重要な情報を発掘するデータマイニングが重要な技術の 1 つになっている．著者らは，Web 上に存在する科学データベースとして，過去から現在までの様々な台風データが収集・研究されている総合データベース，すなわち，デジタル台風に着目している．デジタル台風[1]は国立情報学研究所が公開している台風のデータベースを提供する Web サイトである．

台風とは，熱帯低気圧の中で北西太平洋（赤道より北で東経 180 度より西の領域）または南シナ海に存在し，なおかつ低気圧の中心付近の最大風速（10 分間平均）が 34 ノット（17.2m/s）以上であるものに与えられる名称である．ここでは，日本の気象庁が認めた台風だけに着目し，発生年度と発生順序を組み合わせた台風番号で各台風を表現する．例えば，1990 年度に発生した台風の中で，5 番目に確認されたものは「199005」と表される．

台風は，毎年夏から秋にかけて，日本列島に被害を及ぼしている．気象庁[2]の統計資料によれば，過去 30

年間（1971 年～2000 年），一年間に平均で約 27 個の台風が発生し，そのうち約 11 個が日本から 300km 以内に接近，または上陸している．台風による被害の種類・規模は様々ではあるが，最近の例で言えば，広島市で最大瞬間風速 60.2m/s を記録し，厳島神社の建築物にも被害をもたらした 2004 年の台風 18 号のように，記録的な被害をもたらす台風も存在する．このような事情により，データマイニング技術を用いて，予め，日本に接近する台風の被害予測ができれば，被害を最小限に抑える対策が得られるものと期待される．

本研究では，巨大科学データベースの一つであるデジタル台風のデータベースを用い，データマイニングを行うことで台風の特徴を発見し，その台風の特徴ごとにデータを分類し，台風の予想経路から被害の種類や規模を容易に予測することを可能にする「台風被害予測モデル」の構築手法を提案する．

本稿の構成については，以下のとおりである．まず，2 章では，関連研究について述べる．3 章では，デジタル台風のデータベースについて述べる．4 章では本稿で重要な用語の定義を行う．5 章では，従来手法について述べる．6 章では提案手法について述べ，7 章では評価・考察を行い，8 章では本稿のまとめを行う．

2. 関連研究

デジタル台風のデータベースを対象にした研究としては，(1)画像データを対象とした類似画像検索や画像データマイニング，(2)台風関連ニュース記事のようなテキストデータを対象としたサーチやデータマイニング，(3)アメダス観測データのようなセンサデータを対象としたサーチやデータマイニング，(4)災害データのサーチやデータマイニング，(5)Web サイトにおけるサーチ，(6)統合的情報可視化インタフェースなどがある[3]．(1)から(5)の研究はいずれも単一のデータベースに対するサーチやマイニングの研究が中心になっている．しかしながら，これらの研究には，複数のデータベースに対するサーチやデータマイニングが考慮されていない．(6)の研究では，複数のデータベースに対するサーチの問題を扱っているが，異種データ空間をキーで結合した可視化インタフェースを提案するだけにとどまっている．

本研究では，台風の経路データベースと被害データベースの双方に着目し，台風の予測経路や移動速度から台風の被害予測を行うための台風被害予測モデルの構築手法を提案している．

3. 使用するデータベース

本章では本研究で扱うデータの概要についての説明を行う．本研究では，デジタル台風に格納されている，「台風ごとの被害規模が都道府県別に格納される被害データベース」及び「台風ごとの経路と移動速度が格納される経路データベース」の 2 種類のデータベースを用いる．

ただし，台風被害予測モデルの構築において使用するデータは，詳細な被害情報が残っている 1990 年度以降のデータを利用する．なお，今回は 2004 年度までのものを台風被害予測モデル構築で利用し，後に詳細な被害情報が更新された 2005 年度～2007 年度のデータを，台風被害予測モデルについて評価を行う際に利用する．また，抽出した経路が日本から離れた海上を通過している場合，そのデータから得られる情報が著しく減少するため，本研究では，予め用いる経路のデータベースについて，抽出しようとしている台風の経路は，3.2 節で定義する．指定した緯度・経度の範囲内に侵入するのか否かを検証し，その結果，範囲外のものは除外していく．経路データベースは，日本列島および日本近海を通過する領域だけを使用する．

以後，台風の被害規模が格納されるデータベースを「被害データベース」，台風の特徴や経路が格納されるデータベースを「経路データベース」と呼ぶことにする．以下で両者の詳細について説明する．

3.1. 被害データベース

被害データベースは，日本に接近し日本国内で被害を出した台風を抽出し，台風ごとに，各都道府県の被災・被害の規模をまとめたものである．都道府県全体での被害データベースの総件数は 1636 件であり，被害データベースの属性数は 16 である．表 1 に 16 個の属性を示す．

表 1：被害データベースの属性

番号	被害名称
1	強風（0 か 1）
2	大雨（0 か 1）
3	死者・行方不明（人）
4	負傷者（人）
5	全壊（焼）・流失（棟）
6	半壊（焼）・破損（棟）
7	床上浸水（棟）
8	床下浸水（棟）
9	耕地冠水（ヘクタール）
10	道路損壊（ヶ所）
11	堤防決壊（ヶ所）
12	山崖崩れ（ヶ所）
13	通信施設被害（回線）
14	水産業被害（万円）
15	林業被害（万円）
16	農業被害（万円）

表 1 の番号 1，2 は発生した気象現象のデータであり，被害データは 0 か 1 の値を取り，値が 1 の時その気象現象が発生したことを示す．反対に 0 の時その気象現象が発生していないものとする．また，番号 3～16 までの被害データは発生した被害の実績値を取る.なお，該当する台風の都道府県に一切の被害が発生していない場合は，その都道府県のタプルは欠損値として扱うものとする．

3.2. 経路データベース

経路データベースは，気象機関（気象庁）が発表する「ベストトラックデータ」を利用する．ベストトラックデータとは最終解析結果とも呼ばれ，一定時間ごとの台風の中心位置を専門家が後日に解析してまとめたものである．一定時間は協定世界時 UTC (Universal Time Coordinates)が基準となっており，世界規模の気象観測では UTC の 0 時，12 時に観測が行われる．デジタル台風の経路データは UTC の 0 時，6 時，12 時，18 時に観測されたものであり，日本時間に直すとそれぞれ 9 時，15 時，21 時，3 時（翌日）となる．

本研究で使用する経路データベースは，日本列島が内包される東経 120 度～154 度,北緯 24 度～50 度の領域に限定している．3.1 節の被害データベースから，1990 年度～2004 年度で各都道府県に被害を与えた台風は，合計 107 個存在することが分かった．本研究では，これらの台風について，6 時間ごとに観測された台風の位置情報を知る必要がある．このとき，指定した領域外の位置情報が記録されているものを除くと，使用する経路データベースの総件数は，2814 件となった．なお，経路データベースの属性については，台風の位置情報が実数値で格納される「緯度」,「経度」，及び台風の平均移動速度の情報が実数値で格納される「平均移動速度」の合計 3 種類とする．このとき，実数値で格納されている緯度・経度において，緯度は北から“a”から“z”に，経度は東から“A”から“AH”と置き，小領域ごとに“aA”から“zAH”という 884 個の領域の系列で表現することで，領域内に存在する経路の数値情報を領域の系列に置き換える．

4. 用語の定義

本章では，本研究で用いる用語の定義を行う．

4.1. 台風パターン

台風の特徴は，台風の経路と平均移動速度で表現される．台風の経路は，東西属性と直撃属性の 2 つの属性の値により表現され，平均移動速度は，速度属性の値により表現される．東西属性は，ある基準点から見て，台風が『東寄り』，または『西寄り』のどちらであるかを表現し，直撃属性は，その基準点から見て，台風が『直撃』，または『接近』のどちらであるのかを表現している．速度属性は，台風の平均移動速度が『S（遅い)』，または『M（普通)』，もしくは『F（速い)』のどちらであるかを表現している．従って，台風は，3 つの属性の値を組み合わせた 12 種類の台風パターンで表現される．なお，台風パターンの計算法については，6.3 節で詳しく述べる．

4.2. 気象被害パターン

気象被害パターンは，台風の気象的性質及び台風が及ぼす大まかな被害の種類・規模を表すものである．6.2 節で詳述するが，気象被害パターンの種類は全部で 72 種類存在する.被害データベースのデータをクラスタリングすることにより被害クラスタの集合が抽出される．各被害クラスタ内の台風は，それぞれ類似した台風被害をもたらしており,この 72 種類の気象被害パターンのいずれかで表現される．

4.3. 気象被害パターンの出現数

ある都道府県に現れる台風に着目すると，各台風が帰属する台風パターンおよび気象被害パターンを知ることができる．このとき，ある台風パターンに対して発生した気象被害パターンの件数を，その気象被害パターンの出現数と定義する．なお，気象被害パターンは，各都道府県で異なる．図 1 に例を示す．図中の数値が気象被害パターンの出現数を表している．

台風パターン

気象被害パターン	東						西					
	接近			直撃			接近			直撃		
	S	M	F	S	M	F	S	M	F	S	M	F
パターン1	0	1	0	1	5	2	0	0	4	3	1	0
パターン2	1	0	0	1	1	0	1	0	0	0	0	1
パターン3	0	0	1	0	0	0	0	2	0	0	0	1
パターン4	1	1	0	0	1	0	0	0	1	0	1	0
パターン5	0	0	0	1	2	1	1	1	0	1	0	0

図 1：出現数の例

4.4. 気象被害パターンの出現率

気象被害パターンの出現率とは，気象被害パターンの出現数が各都道府県の中でどれくらいの割合を占めているかを示すものである．気象被害パターンの出現数は，都道府県ごとに台風による被害発生数が異なっているため，気象被害パターンの出現数の値だけでは発生頻度の判断が付けにくいために利用される．気象被害パターンの出現率が高ければ高いほど,すなわち，発生頻度が高いほど，それに該当する都道府県の台風被害予測の信頼度が高いということになる．気象被害パターンの出現率は以下の(1)式で求まる.(1)式の分母の「その都道府県に発生した台風の総数」とは，気象被害パターンの出現率を求めようとしている都道府県

内に限定した台風の総数である．図 1 で言えば，図中の数値の和が分母となる．分子の「出現数」は，4.3 節で述べたように，図中の個々の値である．これを各都道府県で適用する．

$$出現率 = \frac{出現数}{その都道府県に発生した台風の総数} \times 100[\%] \quad (1)$$

4.5. 的中率

的中率とは，後に作成する「台風被害予測モデル」を用いて被害予測したい台風の被害を予測する場合，予測を行った全台風に対して，被害の予測結果がどのくらい的中したかを表す指標である．以下にその式を示す．

$$的中率 = \frac{被害が的中した台風の個数}{被害予測に使用した台風の総数} \times 100[\%] \quad (2)$$

5. 従来手法

本章では，次章のデータ分類・分析で用いた従来手法について説明する．

5.1. クラスタリング手法

クラスタリングとは，データの集まりをデータ間の類似度（或いは非類似度）に従って，いくつかのグループに分ける手法である．クラスタリングの手法は，階層的手法，非階層的手法の 2 種類がある．

5.1.1. 階層的手法

階層的手法とは，与えられたデータセットの各データが 1 つのクラスタとなっている状態を初期状態として，クラスタ間の距離や類似度に基づいて，2 つのクラスタを逐次的に併合してゆく手法を指す．目標のクラスタ数まで併合が行われた時に処理が終了するのだが，通常は 1 つのクラスタになるまで併合を繰り返す．この階層併合的手続きは，AHC（agglomerative hierarchical clustering）と呼ばれる．このときデータの階層構造が得られ，その階層構造はデンドログラム（樹形図）というグラフで表現される．

分類手法は，古典的な手法である郡平均法やウォード法をはじめ，最短距離法，最長距離法などがある．

なお，次章で用いる郡平均法について簡単に説明する．そもそも「最も近い 2 つのクラスタ」を探すためには，2 つのクラスタ同士の距離を決める必要があるのだが，その手法の 1 つとして，2 つのクラスタに属する個体間の全ての組み合わせの距離の平均を利用する方法があり，これを郡平均法と呼ぶ．

大規模かつ高次元なデータに対応するために，BIRCH[5]と呼ばれるような高速にクラスタリングを可能にする方法が数多く提案されている．

5.1.2. 非階層的手法

非階層的手法とは，データの分割の良さを表すある評価関数を設定し，その評価関数に対する最適解（最適な分割）を探索することでクラスタリングを行う手法を指す．MacQueen ら[6]により提案された，クラスタの平均を用い，与えられたクラスタ数 K 個に分類する k-means 法が代表的な手法である．その他に，高次元かつ巨大データベースに対し有用な，Rakesh Agrawal ら[7]による CLIQUE や，Boriana L. Milenova ら[8]による O-Cluster などがある．

5.2. 次元縮約

次元縮約とは，次元の呪い等を回避するために，高次元のベクトル間の位置関係をできるだけ保存した形で，より低次元のベクトルに変換する処理を指す．次元縮約の代表的な手法として，特異値分解に基づく LSI（Latent Semantic Indexing）が知られている．

この，Scott C. Deerwester ら[9]によって提案された LSI では，次元縮約する前に，文書ベクトルの各要素を TF*IDF などで重み付けして，変換しておく必要がある．しかし，文書ベクトルへの重み付けの方法はどれもアドホックであり，この点で，理論的な弱点がある．本研究では，この弱点を回避した pLSI（Probabilistic Latent Semantic Indexing）を利用する．

pLSI は，T.Hofmann[10]によって提案された次元縮約の手法の一つであり，クラスタリングに直接利用することも可能である．pLSI は確率・統計的な枠組みで，LSI と同様の処理を行う．縮約する対象としては，文書ベクトルが想定されているが，文書ベクトルへの重み付けの必要が無く，単純に頻度を要素とするベクトルを使用することができる．文書と単語結び付ける潜在的なクラスを想定した Aspect モデルというモデルを利用し，文書 d と単語 w の出現を潜在的なクラス z を用いて，(3)式が導出される．

$$p(d,w) = \sum_z p(z)p(w|z)p(d|z) \quad (3)$$

K 次に次元縮約する場合は，潜在的なクラスを K 個設定し，データ d に対して，

$$(p(z_1,d), p(z_2,d), \cdots, p(z_K,d)) \quad (4)$$

が縮約されたベクトルとなる．与えられた文書集合が $D=(d_1,d_2,\cdots,d_N)$ であり，D で使われている単語の集合が $W=(w_1,w_2,\cdots,w_M)$ である場合に，(5)式に示す対数尤度関数

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log p(d_i,w_j) \quad (5)$$

を最大化するパラメータを求めれば良い．これを求めるために EM アルゴリズムを用いる．E-step では，

$$Q_{ijk}^{(t+1)} = \frac{p(z_k)^{(t)}\, p(w_j|z_k)^{(t)}\, p(d_i|z_k)^{(t)}}{\sum_{k=1}^{K} p(z_k)^{(t)}\, p(w_j|z_k)^{(t)}\, p(d_i|z_k)^{(t)}} \quad (6)$$

を計算し，M-step では，

$$p(w_j \mid z_k)^{(t)} = \frac{\sum_{i=1}^{N} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}} \quad (7)$$

$$p(z_k)^{(t)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}} \quad (8)$$

$$p(d_i \mid z_k)^{(t)} = \frac{\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) Q_{ijk}^{(t)}} \quad (9)$$

を計算する．E-step と M-step を，値が収束するまで繰り返し行う．なお，t は，繰り返し回数を示している．

実際に EM アルゴリズムで各パラメータを求める際には，初期値（t=0 の時の値）が必要である．

6. 台風被害予測モデルの構築手法

本章では，被害データベースの分類及び経路データベースの分類に基づいて，台風被害予測モデルを構築する手法を提案する．

6.1. 被害データベースのクラスタリング

階層的クラスタリングにより，全部で 1,636 種類ある被害データを最小限の種類のクラスタで表現することができる．ここでは，階層的クラスタリングを行う前に，以下の前処理を行っている．

(1) 全ての属性について属性値を正規化

被害データベースの各属性は，被災・被害の規模を表現しているが，被害データの各属性はそれぞれ異なる尺度で測られているため，ユークリッド距離を用いる場合，各属性の尺度の影響の違いを考慮する必要がある．このため，以下の式を用いて，全ての属性について属性値を 0～1 の間に正規化している．

$$a_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)} \quad (10)$$

(2) 次元縮約

被害データベースの従属性によるクラスタリングの精度低下を防ぐために，必要に応じて pLSI による次元縮約を行う．

図 2 は，次元縮約を行わず正規化だけを施した広島県の被害データを用いて，階層的クラスタリングを行った結果である．次元縮約の効果については，7 章で述べる．

6.2. 気象被害パターンの計算法

図 2 の例で示されたように，ある都道府県におけるクラスタリング結果から，類似した被害をもたらす台風の集まり,すなわち,クラスタの集合が抽出される．以後，抽出された各クラスタを「被害クラスタ」と呼ぶことにする．

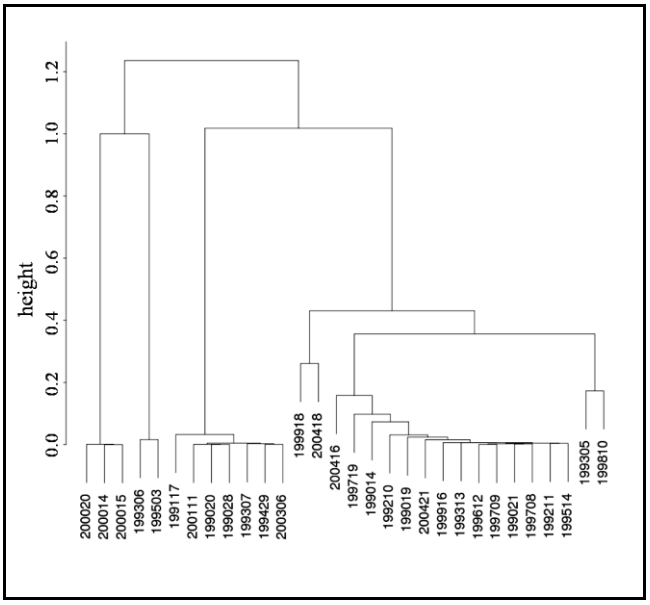

図 2：広島県におけるデンドログラム

この被害クラスタの特徴を表現する気象被害パターンの計算手順は，以下のとおりである．

(1) 被害データを表現する 16 属性（表 1 参照）のそれぞれについて，属性値の平均を計算する．

(2) 上記の 16 属性の平均値を用いて，気象的性質と被害的特徴を決定する．この決定方法については後述する．

(3) 被害クラスタで計算された気象的性質と被害的特徴との組みを気象被害パターンとする．表 2 の作り方は後述するが，表 2 を用いて，気象被害パターンの番号を見つけ出す．

以下では、上記(2)の気象的性質および被害的特徴の計算方法について述べる．

気象的性質については，以下のように決定している．気象的性質の決定は，表 1 の番号 1 及び 2 のそれぞれに対する被害名称の値の記号化により達成する．平均値が 0.5 以上であれば，気象的性質は「強風/大雨が発生する」であると決定する．この決定方法を用いると，発生する台風の気象的性質は,「強風・大雨」,「強風」,「大雨」,「気象的性質無し」の 4 通り存在する．

被害的特徴は，表 1 の番号 3～16 のそれぞれに対する被害名称の値から決定するが，その決定方法の説明をする前に,被害名称の値の記号化について説明する．番号 3～16 の属性のそれぞれに対して，被害名称の平均値を取ることで被害の規模を 4 種類のランクに分類している．被害が発生していない場合は「*D*」，平均値以下であれば「*C*」，平均値を二倍した値以下であれば「*B*」，それ以上の高い値は「*A*」と表現している．

被害的特徴については，以下のように決定している．被害的特徴は，番号 3～16 に対する被害名称の被害規

模が「*D*」のものが半分（7 つ）以上ならば被害が「少ない」，「*A*」のものが 1 つ以上 7 つ未満ならば被害が「ある特定の被害が大きい」，「*A*」のものが半分以上ならば被害が「非常に大きい」，いずれのパターンも当てはまらない場合を被害が「平均的」であると表現することができる．しかしながら，被害的特徴の属性値の「ある特定の被害が大きい」については，実際どのような種類の被害が大きいのか分かり難い．このため，番号 3～16 に対する被害名称を，図 3 に示されるように 4 つのカテゴリーに分割し，カテゴリーの組合せをとることにより，被害的特徴を決定している．この結果，「ある特定の被害が大きい」については 15 通りの表現が可能となり，これに「非常に大きい」，「平均的」，「少ない」を合わせ，被害的特徴を 18 通りで表現している．

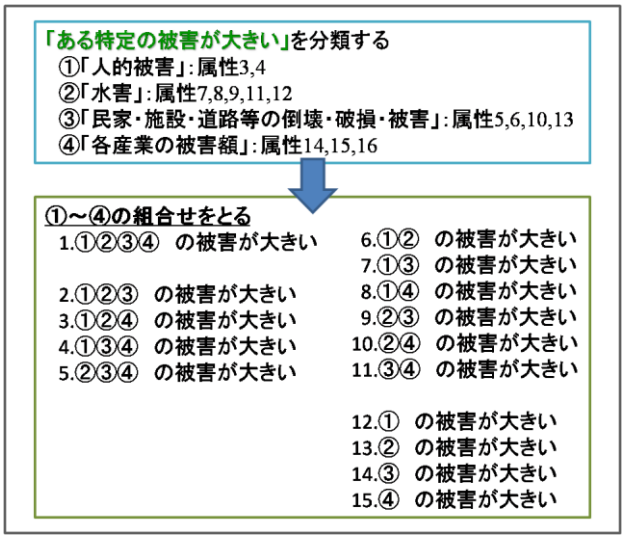

図 3：「ある特定の被害が大きい」の分類

表 2 の作り方については，以下のとおりである．気象的性質の分類結果は 4 通り存在し，被害的特徴の分類結果は 18 通り存在する．従って，これらを組み合せると，合計 72 種類の「気象被害パターン」が存在する．表 2 はこれらの 72 種類の気象被害パターンに番号を付与することにより作られている．

6.3. 台風パターンの計算法

本節では，台風被害予測モデルの構築の際に必要となる，3.2 節で定めた「経路データベース」の分類方法について述べる．まず，台風の移動速度の分類について述べる．過去に発生した各台風の移動速度などにはバラつきが存在し，それらを全て一括りにしてしまうと，情報の信頼度が減少してしまう．そこで，本研究では台風の平均移動速度を，表 3 のように 3 種類のランクに分類する．例えば，図 4 に示す「2004 年度の台風 21 号（200421）」の経路データでは平均移動速度が「20.4km/h」とあるので分類は「M」となる．

表 2：気象被害パターンの種類

気象被害パターンの番号	気象的性質	被害的特徴
1	強風・大雨	非常に大きい
2~16		ある特定の被害が大きい
17		平均的
18		少ない
19	強風	非常に大きい
20~34		ある特定の被害が大きい
35		平均的
36		少ない
37	大雨	非常に大きい
38~52		ある特定の被害が大きい
53		平均的
54		少ない
55	気象的性質無し	非常に大きい
56~70		ある特定の被害が大きい
71		平均的
72		少ない

表 3：平均移動速度の分類

記　号	平均移動速度（km/h）
S	6≦X＜18
M	18≦X＜30
F	30≧X

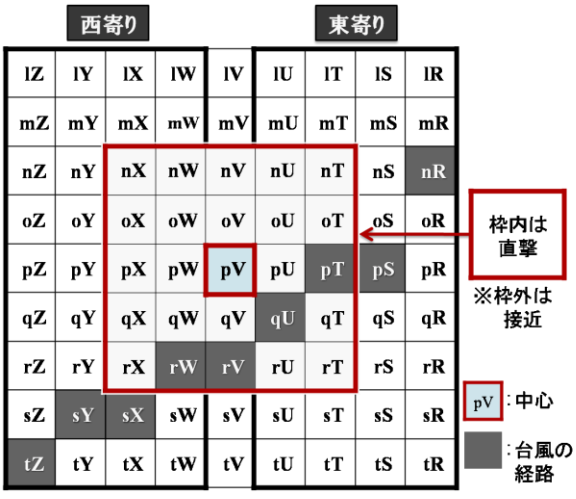

図 4：台風の経路図

次に，各都道府県で台風ごとの経路データの分類を行う．3.2 節で定義した各都道府県の位置情報を領域

の系列で置き換えたデータを使用する．例えば，広島県の経路データの分類を行う場合，図 4 に示したように，広島県の位置情報は“pV”となり，この位置情報を中心にして考える．

15 年間に日本で被害の発生した全ての台風を調査するわけではなく，中心とした都道府県で被害の発生した台風の経路データのみを使用し，そのデータの分類を行う．中心から東にあるものと西にあるものを分類し，台風ごとに「東寄り」の経路なのか，それとも「西寄り」なのか検証する．

6.4. 台風被害予測モデル

被害クラスタごとに，被害クラスタに含まれる台風パターンを調べ，その被害クラスタにどのような台風パターンが何件あるかを数えることにより，台風パターンの出現率を計算する．都道府県ごとに，各被害クラスタに含まれる台風パターンを出現数あるいは出現率とともにまとめたものを台風被害予測データベースと呼ぶ．表 4 は，都道府県として，広島県を選択した場合の台風被害予測データベースの例である．

表 4：広島県における台風被害予測データベース

気象被害パターンの番号	東						西					
	接近			直撃			接近			直撃		
	S	M	F	S	M	F	S	M	F	S	M	F
2	0	0	0	0	0	3	0	0	0	0	9	0
3	0	0	0	0	3	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	3
17	0	0	0	0	3	0	0	0	0	0	3	0
18	0	0	0	3	12	6	0	0	0	6	3	0
35	0	0	0	0	0	0	0	0	0	0	0	3
36	0	3	0	3	3	0	0	3	0	0	6	0
54	0	0	0	0	0	0	0	3	0	0	0	3
72	0	0	0	0	0	0	3	3	0	0	3	0

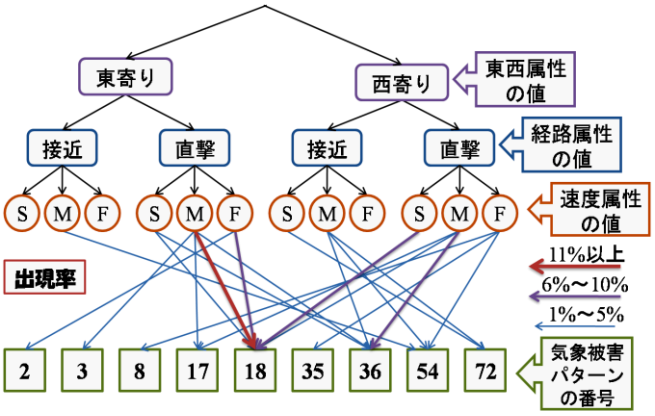

図 5：広島県における台風被害予測モデル

このデータベースを図で表現したものが，台風被害予測モデルである．なお，図 5 に広島県における台風被害予測モデルを示す．例えば「東寄り」の経路の「直撃」コースで，平均移動速度が「M」の台風を被害予測する場合，気象被害パターン 18「強風・大雨の特徴を持つが，被害は少ない」の可能性が最も高いことが分かる．このモデルは，47 都道府県分存在する．

7. 評価・考察

本章では，台風被害予測モデル及び次元縮約の効果について，評価・考察を行う．評価の指標としては，3.3 節で述べた的中率を用いる．なお，台風被害予測モデルの的中率を求めるために，予測モデルの構築で利用しなかった 2005 年度～2007 年度の 3 年分の被害・経路データを用い，予め各都道府県に対し，各台風がどのような経路をとり，表 2 のどの気象被害パターンに該当するか調べておく．そして，これらの台風を予測モデルに入力して求まった，可能性が最も高い気象被害パターン番号と，予め調べておいた気象被害パターン番号が一致するか否かを調査し，(2)式を用いて的中率を求める．次元縮約を施す場合については，次元数ごとに，クラスタ数を 5～20 個に変化させて的中率の変化を見る．なお，pLSI 使用時の各パラメータの初期値については，各次元とも適当に何通りか決め，的中率が最も良いと思われるものを採用する．繰り返し回数は，値の収束具合から判断し，100 回とした．

結果としては，クラスタ数が多い場合に的中率が良い傾向となった．より細かいグルーピングが，的中率の上昇に繋がったと考えられる．また，縮約する次元数について的中率から判断すると，次元数が 4 次元以上になると，的中率はほぼ横ばいになった．ここでは，できるだけ次元数を減らし，無駄な属性を除去することが目的であるので，的中率から判断すると，4 次元が適当であると考えられる．参考までに表 5 に，2005 年度及び 2007 年度のデータを用いた場合の，クラスタ数 5～20 における的中率の平均値と最大値をまとめた．2006 年度のデータ件数は，41 件で，他の年（2005 年度は 74 件，2007 年度は 87 件）に比べかなり少なく，信頼性が低いと考えられ，ここでは省略する．

表 5：各次元での的中率の平均値・最大値

次元数	2005 年度		2007 年度	
	平均値	最大値	平均値	最大値
2 次元	28.69%	39%	25.38%	35%
3 次元	30.69%	45%	24.69%	37%
4 次元	29.38%	47%	25.56%	40%
5 次元	29.50%	44%	23.44%	36%
6 次元	32.56%	50%	25.31%	37%
7 次元	32.50%	50%	23.63%	37%
10 次元	30.56%	50%	21.81%	37%
縮約無し	27.75%	37%	20.94%	35%

8. まとめ

本研究では，pLSI に基づく次元縮約と階層的クラスタリングを用いて，簡易的な台風被害予測被害モデルの構築方法を提案し，その結果，大まかな被害予測が可能となった．今後の課題を以下に示す．

階層的クラスタリングでは，郡平均法を用いたが，ウォード法や最長距離法などの，他の手法についても検証する必要がある．他の手法を用いることにより，的中率が改善される可能性がある．

また，本研究では，被害データベースの従属性を考慮し，pLSI を用いた次元縮約をクラスタリングの前処理として行ったが，pLSI 以外にも LSI を用いることが可能であり，両者の精度について比較検討を行う必要がある．なお，pLSI においては，局所最適性の問題を有し，この点の解決も必要である．

それから，領域分割における記号化の際に発生する曖昧性の解消や，経路分類のパターンを増やすことにより正確な台風被害の予測が可能であると考えられる．

謝　辞

本研究の一部は，日本学術振興会，科学研究費補助金（基盤研究（C），課題番号：20500137）の支援により行われた．

文　献

[1] デジタル台風:
http://agora.ex.nii.ac.jp/digital-typhoon/

[2] 気象庁:
http://www.jma.go.jp/jma/index.html

[3] 北本 朝展:"デジタル台風: 大規模時系列データのマイニングとサーチ", 電子情報通信学会 データ工学研究専門委員会 第二種研究会チュートリアル(招待講演), pp.21-49, 2007 年 11 月.

[4] Trevor Hastie, Robert Tibshirani, Jerome Friedman: 14.3 Cluster Analysis, The Elements of Statistical Learning(2nd ed.), Springer-Verlag, pp.501-528, 2009.

[5] Tian Zhang, Raghu Ramakrishnan, Miron Livny: BIRCH: An Efficient Data Clustering Method for Very Large Databases. Proc. of the ACM SIGMOD Conference on Management of Data, ACM Press, pp.103-114, 1996.

[6] MacQueen, J. B.: Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol.1, University of California Press, pp.281-297, 1967.

[7] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan: Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Proc. of the ACM SIGMOD Conference on Management of Data, ACM Press, pp.94-105, 1998.

[8] Boriana L. Milenova, Marcos M. Campos: O-Cluster: Scalable Clustering of Large High Dimensional Data Sets, Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), pp.290-297, 2002.

[9] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Richard A. Harshman: Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science (JASIS), 41(6), pp.391-407, 1990.

[10] Thomas Hofmann: Probabilistic Latent Semantic Indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp.50 – 57, 1999.