サポートベクターマシンを用いた気圧配置の自動分類

木村　広希† 　川島　英之†,†† 　北川　博之†,††

† 筑波大学システム情報工学研究科　〒 305–8573 茨城県つくば市天王台 1–1–1
†† 筑波大学計算科学研究センター　〒 305–8573 茨城県つくば市天王台 1–1–1
E-mail: †hkimu@kde.cs.tsukuba.ac.jp, ††{kawasima,kitagawa}@cs.tsukuba.ac.jp

あらまし　気象分野では，西高東低冬型などの特定の気圧配置に着目して研究を行うことがある．このとき研究者は，数十年分の過去のデータから特定の気圧配置の事例を集めるために，天気図を一枚ずつ目視で判別しているのが現状である．この目視による判別は，必要とするデータの数が多いほど過酷な作業となる．この問題を解決するため，本稿では，気圧配置に対してサポートベクターマシンを用いて自動で判別することを提案する．分類対象の気圧配置は，西高東低冬型，気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型とし，JRA-25 データを用いて分類実験を行った．また，実験において最良の結果を示した分類器を用いて 1979 年〜2006 年のデータを分類し，その分類結果をもとに検索システムのプロトタイプを開発し，アンケート調査による評価を行った．
キーワード　データマイニング，サポートベクターマシン，気象学，気圧配置，特徴抽出

Automatic Classification of Pressure Patterns by Support Vector Machine

Hiroki KIMURA†, Hideyuki KAWASHIMA†,††, and Hiroyuki KITAGAWA†,††

† Graduate School of Systems and Information Engineering, University of Tsukuba　Tennodai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
†† Center for Computational Sciences, University of Tsukuba　Tennodai 1–1–1, Tsukuba, Ibaraki, 305–8573 Japan
E-mail: †hkimu@kde.cs.tsukuba.ac.jp, ††{kawasima,kitagawa}@cs.tsukuba.ac.jp

Abstract　In the field of meteorology, when researchers need the data that has a specific pressure pattern, for example "Low in West and High in East" or "High in South and Low in North", they have to see huge data by their own eyes in order to judge the data has the specific pressure pattern. To deal with this problem, we propose an automatic classification method by Support Vector Machine. In this study, we tried to detect the winter type, the trough type, the migratory anticyclone type, the flont type, the summer type and the typhoon type. Using the classifier that obtained the highest F-measure in the result, we developed a prototype of pressure pattern search system. In the result of the inquiry survey, we verified the availability of the automatic classification.
Key words　Data Mining, Support Vector Machine, Meteorology, Pressure Pattern

1. はじめに

近年自然科学分野では，蓄積されたデータ資源を有効に活用した様々な研究が行われている．気象分野においても，再解析データなどの格子点データの普及が進み，計算機を用いた大規模データの統計解析や，数値シミュレーションなどが行われるようになった．現在数値予報によって日々の天気予報が行われているように，気象現象の予測の需要は高い．

ここで，気象現象を予測するための情報の一つに，気圧配置がある．気圧配置とは高気圧と低気圧の分布を示し，よく知られている例として，西高東低冬型の気圧配置になると，北から冷たい風が吹き込み，冷え込むというものがある．気象分野での気圧配置に関する研究では，西高東低冬型や南高北低夏型などの，ある特徴をもつ気圧配置の事例を多数必要とすることがある．膨大な量の過去の事例が蓄積されているが，それぞれの事例が条件を満たすかどうかの判別は，現状では多くの場合，天気図の目視によって行われている．必要な事例の数が多いほど，この目視による判別という作業は困難なものとなる．

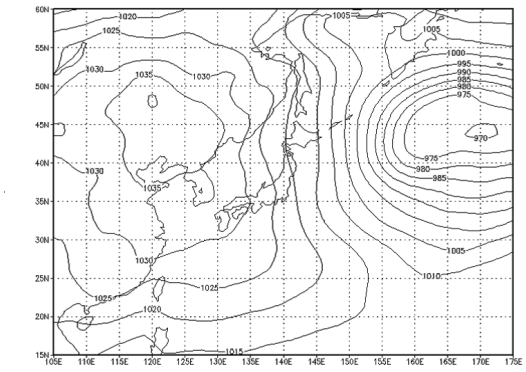


図 1　対象領域の海面更正気圧

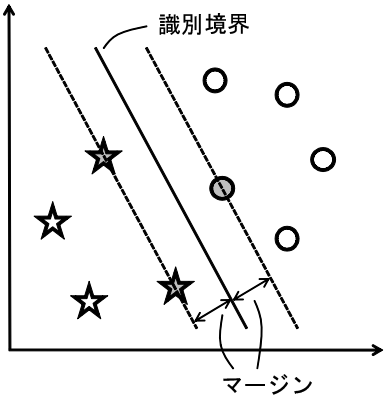


図 2　SVM による線形分離

そこで，本稿では気圧配置の分類に対してサポートベクターマシン (以下 SVM) を適用することにより，日々の気圧配置を自動的に分類する方法を提案する．実験では，対象データの期間を 1981〜2000 年として分類実験と評価を行った．学習データとして，気象データには JRA-25 [1] のデータを，気圧配置を示すデータには [2] による分類を用い，SVM の計算には，SVM のツールである TinySVM [4] を用いた．また，実験で得られた最良の分類器を用いて 1979 年〜2006 年のデータを分類し，その分類結果をもとに検索システムを開発した．

本稿の構成は次の通りである．まず，2 節で本研究の分類対象の気圧配置と対象データについて説明し，3 節で提案手法を述べる．4 節で実験について，5 節で検索システムについて説明し，最後にまとめと今後の課題を述べる．

2.　分類対象の気圧配置と学習データ

2.1　分類対象の気圧配置

気象学の気候の分野において代表的な文献である [2] では，基本的な気圧配置を 6 種類(注1)に分類している．また，6 種類の型のいずれか一つの型に定まらず，複数の型を持つものを移行型もしくは複合型に分類している．本研究では，この [2] による 6 種類の気圧配置を分類対象とする．

ここで，気圧配置に対して 6 クラス分類を行うと，移行型と複合型のデータ，すなわち複数の型の特徴を持つデータが適切に分類できないと考えられる．そのため本研究では，特徴を ”持つ ”と ”持たない ”の 2 クラス分類器を，気圧配置ごとに作成して分類を行う．先行研究 [3] では，6 種類の気圧配置のうち，西高東低冬型の分類を行った．本研究では，西高東低冬型に加えて，残りの 5 種類の気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型の分類を行う．

2.2　学習データ

本研究では，SVM により分類を行うため，学習データとして，日々の数値データとそれぞれの日のクラスラベル（気圧配置）を示すデータが必要となる．日々の数値データには JRA-25 [1] のデータを用い，1981 年から 2000 年の日本時間 9 時の日ごとのデータを対象とする．また，それぞれの数値データのクラスラベルを示すデータには, [2] による気圧配置の分類を用いる. [2] には，対象とする数値データと同時刻の気圧配置の分類が掲載されている．

ここで，JRA-25 とは，気象庁と電力中央研究所による長期再解析を意味する．JRA-25 のデータは，気圧や気温など様々な要素を持つ，南北 1.25 度・東西 1.25 度メッシュの全球データである．本研究では，日本周辺の地域である北緯 15〜60 度・東経 105〜175 度の数値データ（37 × 57 = 2109 点）を対象とする．図 1 に対象領域における海面更正気圧分布を示す.

3.　提 案 手 法

3.1　サポートベクターマシン (SVM)

本研究では 2 クラス分類を行うため，分類手法に SVM を用いる．SVM は，文字認識や音声認識などに幅広く用いられているものである．SVM は，与えられた学習データの中の，サポートベクトル (識別境界近傍に位置する学習データ) と識別境界との距離であるマージンを最大化するように識別境界を構築し 2 クラス分類を行う．図 2 は SVM による線形分離の概念を示す．図 2 において，○がクラス 1 の学習データ，☆がクラス 2 の学習データを示す．各クラスの網掛けされている学習データがサポートベクトルとなり，サポートベクトルと識別境界の距離であるマージンが最大となるように，識別境界を構築する．ここで，本研究では 2109 点数値データを 2109 次元ベクトルとして扱う．

図 2 のような線形分離ができない場合には，非線形分離が行われる．図 3 に非線形分離の概念を示す．これは，非線形な写像 Φ を用いて入力空間をより高次の空間に写像し，写像先の空間で線形分離を行うことで分離を容易にすることを意味する [5]．写像 Φ を用いた計算は，本来であれば元の次元より高次元でのベクトル計算を行う必要があるが，カーネルトリックという手法を用いることで高次元でのベクトル計算を避けることが可能となる．カーネルトリックとは，写像 Φ に関する計算 $\Phi(\boldsymbol{x}) \cdot \Phi(\boldsymbol{x'})$ を，カーネル関数 $K(\boldsymbol{x}, \boldsymbol{x'})$ を用いて置き換えることで写像 Φ の求解を避け，カーネル関数のみの計算に変換することを示す [5]．この計算により，マージンを最大とする識別

(注1)：西高東低冬型，気圧の谷型 (a〜d)，移動性高気圧型 (a〜d)，前線型 (a〜b)，南高北低夏型，台風型 (a〜c)

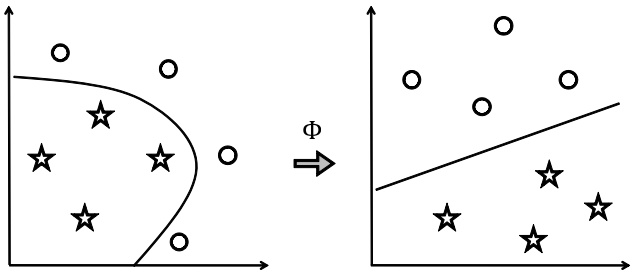


図 3　非線形分離

境界が定まり，識別関数が導出される．

カーネル関数は複数あるが，本研究では，代表的なカーネルの一つである RBF（Radial Basis Function）カーネルを用いた．以下に，RBF カーネルを示す．

$$K(\boldsymbol{x}, \boldsymbol{x'}) = exp(-\delta \times \mid \boldsymbol{x} - \boldsymbol{x'} \mid^2)$$

$\boldsymbol{x}, \boldsymbol{x'}$ はベクトルを，δ はパラメータを表す．RBF カーネルを用いる際には，パラメータ δ に適切な値を与えることが必要である．

以下に，SVM を用いた分類の実行手順を示す．

1. 1 日分の数値データ（ベクトル）と気圧配置（クラスラベル）を 1 組のデータとし，このデータの日数分の集合を学習データとする．
2. SVM により，学習データから分類器（識別関数）を生成する．
3. 生成した分類器（識別関数）を用いて，テストデータを分類する．

3.2　学習データの生成

分類対象の気圧配置は，それぞれ特徴が異なるため，複数の手法で学習データ用のベクトルデータを生成する．以下に，ベクトルデータの生成手法を示す．

手法 1　気圧配置を分類する際に，最も特徴を表すと考えられる海面更正気圧の数値データから，1 件 2109 次元のベクトルデータを生成する．

手法 2　前線を解析するときに用いられる相当温位の数値データから，1 件 2109 次元のベクトルデータを生成する (前線型)．

手法 3　海面更正気圧に加えて気温の数値データから 1 件 $2109 \times 2 = 4218$ 次元のベクトルデータを生成する．

手法 4　海面更正気圧に加えて風速の数値データから 1 件 $2109 \times 2 = 4218$ 次元のベクトルデータを生成する．

手法 5　海面更正気圧データの領域を変化させてベクトルデータを生成する．緯度と経度の範囲を以下のように設定し，すべての組合せで計 20 通りのベクトルデータを生成する．

- 緯度：北緯 15〜60 度，20〜55 度，25〜50 度，30〜45 度．
- 経度：東経 105〜175 度，110〜170 度，115〜165 度，120〜160 度，125〜155 度．

手法 6　手法 5. と同様に相当温位データの領域を変化させてベクトルデータを生成する (前線型)．

手法 7　分類の際に時間軸を考慮するため，対象の事例とその前後の時刻の事例の海面更正気圧の数値データから，1 件 $2109 \times 3 = 6327$ 次元のベクトルデータを生成する．

手法 8　手法 7. と同様に時間変化を考慮するため，対象の事例とその前後の時刻の事例の海面更正気圧の差分値から，1 件 $2109 \times 2 = 4218$ 次元のベクトルデータを生成する．

以上の手法でベクトルデータを生成し，SVM により分類を行い，分類結果を比較する．

表 1　分類結果の正誤評価

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

4. 実　　験

4.1　実験データと評価

実験データには，2.2 節に示す 1981〜2000 年の数値データを用い，[2] の分類により正例と負例を与えた．また，SVM の計算には TinySVM [4] を用いた．なお，学習データのベクトルデータは，3.2 節に基づいて生成した．

1981〜1990 年と 1991〜2000 年の 2 グループに分け，一方を学習データ，もう一方をテストデータとして分類実験を行い，適合率，再現率，F 値を用いて評価を行った．表 1 に分類結果の正誤評価を示す．また，表 1 を用いて，適合率，再現率，F 値の計算式を以下に示す．

$$\text{適合率} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

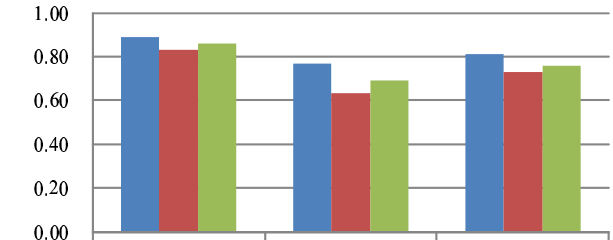
$$\text{再現率} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\text{F 値} = \frac{2\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FN})}$$

4.2　実験結果と考察

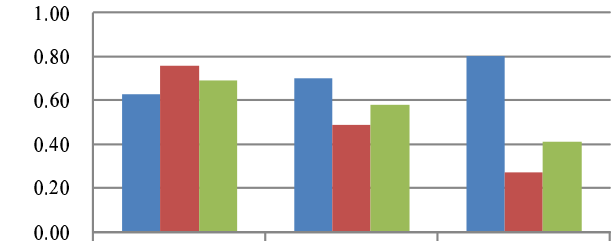
図 4 に西高東低冬型，気圧の谷型，移動性高気圧型，図 5 に前線型，南高北低夏型，台風型の実験結果を示す．この実験結果は，3.2 節に基づいて生成した学習データを用いて行った実験の中で，F 値が最も高い結果を示す．

西高東低冬型は分類結果は，3.2 節の手法 5 の北緯 30〜45 度，東経 115〜165 度の領域で最良の結果が得られ，適合率，再現率，F 値がそれぞれ，0.89，0.83，0.86 となった．同様に，気圧の谷型の分類結果は，手法 5 の北緯 25〜50 度，東経 125〜155 度の領域で，適合率，再現率，F 値がそれぞれ，0.77，0.63，



	西高東低冬型	気圧の谷型	移動性高気圧型
適合率	0.89	0.77	0.81
再現率	0.83	0.63	0.73
F値	0.86	0.69	0.76

図 4　西高東低冬型，気圧の谷型，移動性高気圧型の実験結果



	前線型	南高北低夏型	台風型
適合率	0.63	0.70	0.80
再現率	0.76	0.49	0.27
F値	0.69	0.58	0.41

図 5　前線型，南高北低夏型，台風型の実験結果

0.69 となり，移動性高気圧型の分類結果は，手法 5 の北緯 25〜50 度，東経 115〜165 度の領域で，適合率，再現率，F 値がそれぞれ，0.81，0.73，0.76 となった．また，前線型の分類結果は，手法 6 の北緯 15〜60 度，東経 115〜165 度の領域で，適合率，再現率，F 値がそれぞれ，0.63，0.76，0.69 となった．さらに，南高北低夏型の分類結果は，手法 5 の北緯 30〜45 度，東経 105〜175 度の領域で，適合率，再現率，F 値がそれぞれ，0.70，0.49，0.58 となり，台風型の分類結果は，手法 5 の北緯 25〜50 度，東経 125〜155 度の領域で，適合率，再現率，F 値がそれぞれ，0.80，0.27，0.41 となった．

実験結果から，3.2 節のベクトルデータの生成手法の中では，手法 5 と手法 6 において，数値データの領域を変化させる手法が最もよい F 値が得られることがわかった．手法 1，手法 2 の F 値と比べると，西高東低冬型は 0.03，気圧の谷型は 0.05，移動性高気圧型は 0.01，前線型は 0.01，南高北低夏型は 0.1，台風型は 0.12 高い結果となった．これより，手法 1 と手法 2 と比べて，南高北低夏型と台風型はより特徴を示す領域で分類できており，その他の型は精度を維持しながら不要なデータ点を省けていると考えられる．また，手法 3 と手法 4 では，海面更正気圧に加えて，それぞれ気温と風速の数値データからベクトルデータを生成し，手法 7 と手法 8 では，時間変化を考慮してベクトルデータを生成したが，適切に特徴を抽出できなかったと考えられる．

実験結果に着目すると，全体を通して，それぞれの気圧配置

図 6　検索システムの外観

の分類において移行型と複合型の事例の誤分類が見られた．これは，移行型と複合型の事例が，もう一方の型の特徴の影響を受けて正しく分類されなかったと考えられる．

気圧配置ごとに結果を比較すると，西高東低冬型と移動性高気圧型は比較的高い精度が得られていることから，海面更正気圧から特徴が得られていると考えられる．一方，前線型については適合率が低い結果となった．前線型は相当温位データを用いているが適合率の低さから，適切に特徴を抽出できていないと考えられる．

また，気圧の谷型，南高北低夏型，台風型は，再現率がそれぞれ低い値となった．気圧の谷型については，西高東低冬型の事例に現れている低気圧の誤検出や，移動性高気圧型などとの複合型の見逃しによる誤分類が見られた．南高北低夏型については，西高東低冬型と比べて高気圧と低気圧の位置関係のずれが大きい事例が多く，学習の際に特徴を得られていないと考えられる．また，事例の数が少ないため，見逃しの数が再現率の低さに大きく影響を与えていると考えられる．同様に，台風型についても事例数が少ないため，見逃しの数が再現率を表していると考えられる．台風型については，台風があるにも関わらず，[2] の分類では台風型には分類されていない事例が見られた．これは，分類する人の主観によって重要視する気圧配置の特徴が異なり，分類結果が変わることを示している．

本実験では，学習データに与える数値データを様々な手法で生成したが，いくつかの気圧配置型においては，特徴抽出が不十分であった．これより，数値データを変化させるだけでなく，学習データとして用いるデータを厳選することが考えられる．学習データから，過学習による誤分類の原因になるデータを検出し，取り除くことで分類精度が向上する可能性がある．

5.　検索システムの開発とその評価

5.1　検索システムの開発

分類実験において，最も高い F 値を示した分類器を用いて 1979 年〜2006 年 (日本時間 3，9，15，21 時) のデータをすべて分類し，その分類結果をもとに検索システムのプロトタイプ

図 7　検索結果表示画面

を開発した．開発環境は PHP と MySQL である．図 6 に検索システムの外観を，以下に検索条件を示す．

- 気圧配置：　西高東低冬型，気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型
- 期間 (年と月)：　1979 年～2006 年，1 月～12 月
- 時刻：　日本時間 3，9，15，21 時
- 確信度 (SVM による分類の際の識別境界の基準値)：-1～1（0.2 刻み）

以上の検索条件を満たす事例の日付・時刻を可視画像 (海面更正気圧と相当温位) と共に検索結果として提示する．図 7 に検索結果の表示画面を示す．

ここで，検索条件の SVM による分類の際の識別境界の基準値について以下で述べる．

SVM による分類結果は 1 クラスと-1 クラスの 2 値だが，実際には識別関数の出力は実数であり，識別境界の 0 を基準として 0 以上を 1 クラスに，0 未満を-1 クラスに分類している．そこで，本システムではこの識別境界の基準値を検索条件のパラメータに含め，任意に値を変化させることにより，検索条件に柔軟性を持たせる．

図 8～図 13 に，4.2 節の実験結果において，識別境界の基準値を-1～1 に変化させた場合の適合率と再現率の推移を示す．図 8～図 13 から，識別境界の基準値を大きくすると適合率が上昇し再現率が低下していることがわかる．これは，1 クラスに分類する条件を厳しくすることにより，分類結果の正確性が高くなっていることを示している．一方，識別境界の基準値を小さくすると適合率が低下し再現率が上昇している．これは，1 クラスに分類する条件を緩くすることにより，分類結果の網羅性が高くなっていることを示している．これより，検索条件で識別境界の基準値を設定することにより，検索結果を柔軟に出力できると考えられる．

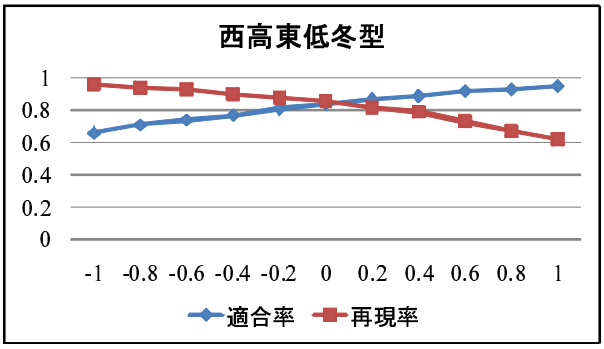


図 8　適合率と再現率の推移 (西高東低冬型)

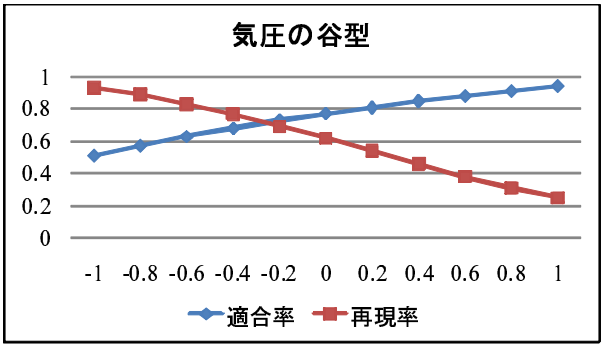


図 9　適合率と再現率の推移 (気圧の谷型)

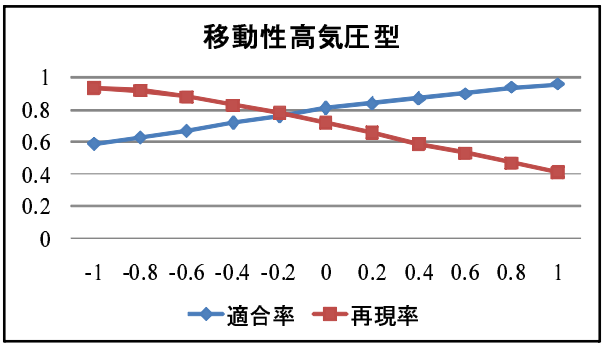


図 10　適合率と再現率の推移 (移動性高気圧型)

5.2　アンケートによる検索システムの評価

検索システムの有用性を検証するために，気象分野の学生を対象に以下のアンケート調査を行った．

- 問 1：西高東低冬型，気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型のそれぞれの検索結果の満足度はどうだったか．
- 問 2：ある気圧配置の事例の収集に有益であると思うか．

図 14 に質問 1，図 15 に質問 2 の回答結果を示す．

問 1 の結果より，それぞれの気圧配置についておおむね高い満足度が得られた．また，全体として，検索条件の識別境界の基準値の高さが検索結果の正確性に反映されているという意見が得られた．

一方で，気圧の谷型では，基準値を高くしても，より顕著な特徴を持つ事例が得られないという意見も得られ，比較的低い

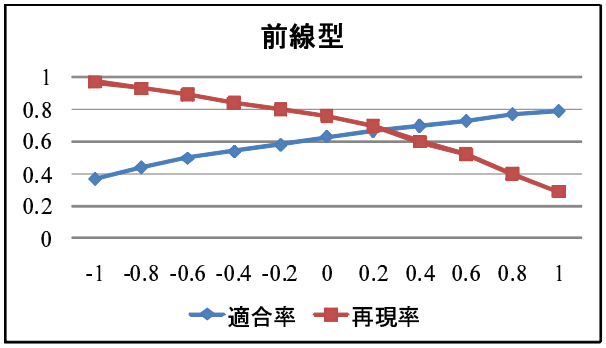


図 11　適合率と再現率の推移 (前線型)

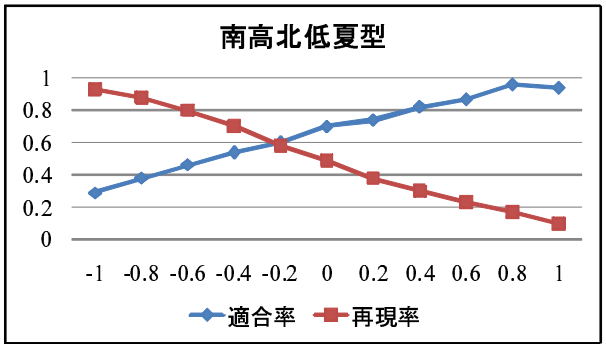


図 12　適合率と再現率の推移 (南高北低夏型)

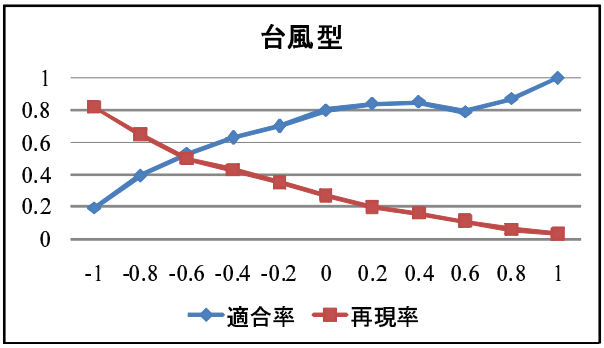


図 13　適合率と再現率の推移 (台風型)

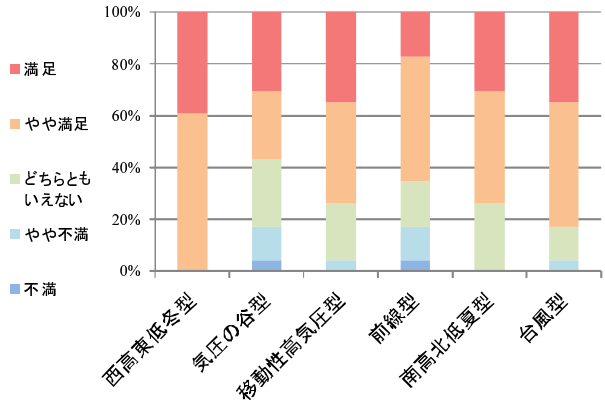


図 14　検索結果の満足度

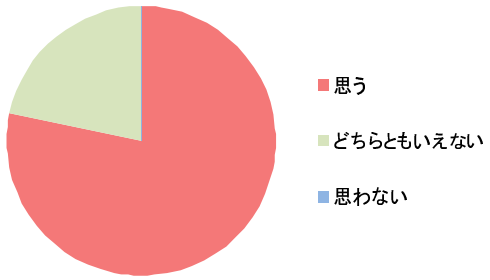


図 15　ある気圧配置の事例の収集に有益であるか

満足度となった．これより，気圧の谷型については，識別関数の出力結果の値の大きさ（識別境界からの距離）と特徴の強さが相関を持たず，特徴を正確に抽出できていない可能性があると考えられる．また，前線型については実験結果と同様に，比較的満足度が低い結果となった．

西高東低冬型と移動性高気圧型は 4.2 節の実験結果と同様に，高い満足度が得られている．また，南高北低夏型と台風型については，実験では F 値が低い結果であったのに対して，比較的高い満足度が得られている．これより，適合率がそれぞれ 0.70, 0.80 であるため，ある程度の正確性において検索結果を提示できていると考えられる．

問 1 の結果より，4.2 節の実験結果の F 値があまり高くなかった気圧配置についても比較的高い満足度が得られた．これは，適合率が示す検索結果の正確性が，ある程度高かったと考えられる．しかし，今回の調査では，再現率が示す検索結果の網羅性は評価できていない可能性があるため，今後検討する必要がある．

問 2 の結果より，高い割合で有益であるという意見が得られた．これにより，SVM を用いて気圧配置を自動分類した結果をもとに，検索システムで検索結果として提示することは有用性があると考えられる．

5.3　ユーザーによる分類結果へのフィードバック

気圧配置の分類は，主観に影響されることがあるため，多くの人の分類結果を集約できれば，より信頼度が高くなると考えられる．そのため，本システムでは，検索結果として出力されたそれぞれの事例に対して，分類結果としてふさわしいか否かをユーザーが投票できる機能を備えている．この投票により集められたユーザーの分類結果を学習データとして利用し，SVM によって分類器を作ることで，よりよい精度の分類結果が得られると期待される．

6.　まとめと今後の課題

本研究では，気象分野の研究における，目視での気圧配置の判別の負担を軽減するため，SVM を用いた気圧配置の自動分類を提案した．分類対象の気圧配置は，西高東低冬型，気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型とした．実験では，SVM の正例と負例の設定に [2] の気圧配置分類を利用し，数値データには JRA-25 のデータを用いた．

実験結果から，西高東低冬型，気圧の谷型，移動性高気圧型，前線型，南高北低夏型，台風型において，それぞれ最良で 0.86, 0.69, 0.76, 0.69, 0.58, 0.41 の F 値が得られた．また，実験

で最良の分類結果を示した分類器で 1979 年〜2006 年のデータを分類し，その分類結果をもとに検索システムのプロトタイプを開発した．さらに，検索システムのアンケート調査による評価を行い，SVM を用いた気圧配置の自動分類の有用性を検証した．

今後の課題は，分類精度の向上と検索システムの実用化である．分類精度の向上としては，学習データに用いるデータの与え方を変えることが考えられる．気圧配置の分類は個人の主観が反映されるため，学習データにミスラベルデータまたは，特徴が弱いデータが含まれている可能性がある．学習の際に，このミスラベルデータや特徴が弱いデータの過学習によって，分類結果の誤分類が生じると考えられる．そのため，学習データから誤分類の原因となるデータを検出・除去したデータを用いて，分類器を生成することを検討している．

また，検索システムの実用化も今後の課題である．ユーザーからの投票が増えれば，より有益な分類結果が得られると考えられる．

謝辞　本研究の一部は科学研究費補助金（# 18650018，# 20240010）と未踏 IT 人材発掘・育成事業の支援による．

文　献

[1] Onogi，K．and Coauthors．”JRA-25： Japanese 25-year re-analysis project-progress and status．” Quart．J．Roy．Meteor．Soc．，131，3259-3268．

[2] 吉野正敏．”日本の気候　最新データでメカニズムを考える”．二宮書店，2002．

[3] 木村広希，川島英之，北川博之．”気象データにおける特徴的気圧配置の自動分類” DEWS2008, 2008．

[4] TinySVM．
http://chasen.org/˜taku/software/TinySVM/

[5] 津田宏治．”サポートベクターマシンとは何か”．電子情報通信学会誌，83：460-466，2000．