データマイニングによる気圧配置の分類

木村 広希 † 川島 英之 †† 北川 博之 ††
† 筑波大学第三学群情報学類 †† 筑波大学大学院システム情報工学研究科

1 はじめに

気圧パタンには様々なものがあり, 文献 [2] においてそれらは 15 種類に分類されている. 気象研究者が研究を進める際に,「西高東低」や「南高北低」などの, ある特徴をもつ気圧パタンを求めることがある. これを実現するには, 目視以外の手法は存在しないのが現状である.

人間の目視でしかなされていない分類を機械により自動化することを目指し, 本研究ではデータマイニングによる「西高東低」(図 1) 分類手法の開発を研究課題と設定する. 気象データには JRA25 データを用い, 対象期間を 1981 年〜2000 年とする. 研究手法としては,Support Vector Machine(SVM) を利用した手法を提案する. 我々の知る限り, 本研究は気圧パタン分類を計算機により自動化する最初の研究である.

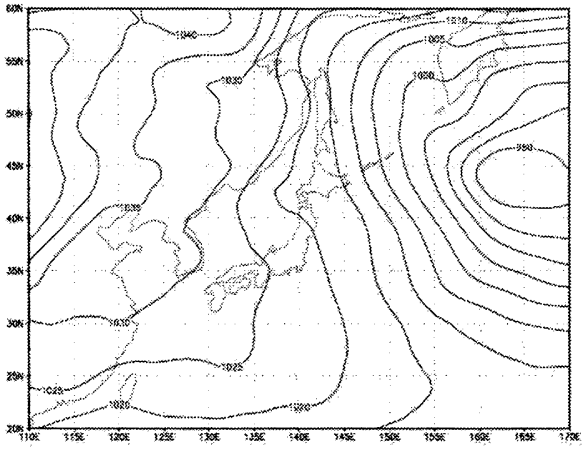

図 1: 西高東低型の気圧図

2 対象データと正例

2.1 対象データ

対象となる気象データには GRIB データ形式である JRA25 を用いる.JRA25 とは GAME 再解析のあと, 日本の気象庁と電力中央研究所で約 25 年 (実際には 1979–2004 年の 26 年間) の長期再解析 (過去にさかのぼったデータ, 一定の同化システムによるデータ同化) を行ったものである.JRA25 データは筑波大学計算科学研究センターによりアーカイブされており,6 時間毎のデータを取得可能である.

JRA25 データの形式は GRIB(GRId Binary) と呼ばれる.GRIB データ形式は大量の格子点データの高圧縮パッケージ法として WMO(The World Meteorological Organization) が 1985 年に提唱したものである.GRIB の他に有名なデータ形式として NetCDF があるがその実現方法は全く異なっており, 扱うデータに制限が多い. そのため本研究では GRIB データ形式を用いる.

JRA25 データは気圧データ以外にも風速などの他種類のデータを含む. そこで各気象データから必要な気圧データのみを抜き出すために wgrib と呼ばれるプログラムを使用した. これにより南北 1.25 度・東西 1.25 度メッシュにおける各格子点の数値データを抜き出した. また, 実験では日本周辺の北緯 15〜60 度・東経 105〜175 度 (37 × 57=2109 点) を対象とした.

2.2 西高東低の正例

西高東低の正例は文献 [2] にある「西高東低冬型」から得た. 文献 [2] では基本的な気圧パタンを 15 種類†としている. また,15 種類のどの型にも属さないものを移行型もしくは複合型に分類しており, これらは 15 種類のパタンのうちの複数の型を持つ. 本研究では, 文献 [2] に記載されている 1981〜2000 年の 20 年間分のデータを正例・負例の判別に利用する.

3 提案

3.1 Support Vector Machine

SVM(Support Vector Machine) は 2 クラスの分類を行う機械学習手法の一種である. 与えられた訓練点のなかでサポートベクトルと呼ばれるクラス境界近傍に位置する訓練点と識別面との距離であるマージンを最大化するように分離超平面を構築しクラス分類を行う. 線形分離が難しい際には, カーネルトリックによって入力空間をより高次の特徴空間に写像し, そこで線形分離を行うことで分離を容易化できる.

各格子点の気圧の数値を 2109 次元のベクトルとみ

Classification of Barometric Pressure Patterns by Data Mining
† Hiroki KIMURA(hkimu@kde.cs.tsukuba.ac.jp)
†† Hideyuki KAWASHIMA(kawasima@cs.tsukuba.ac.jp)
†† Hiroyuki KITAGAWA(kitagawa@kde.cs.tsukuba.ac.jp)
College of Information Science, University of Tsukuba, University of Tsukuba (†)
Graduated School of Information and Systems Engineering, University of Tsukuba (††)
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

† 西高東低冬型, 気圧の谷型 a〜d, 移動性高気圧 a〜d, 前線型 a〜b, 南高北低夏型, 台風型

中間例の有無	カーネル	precision(S)	recall(S)	precision(¬ S)	recall(¬ S)	正解率
両方有	線形	87.2	80.9	94.2	96.3	92.0
	非線形	90.0	83.0	94.9	97.2	93.4
学習のみ有	線形	76.8	89.7	98.5	96.2	95.2
	非線形	80.5	91.7	98.8	97.0	96.2
検証のみ有	線形	91.8	74.3	91.7	97.7	90.9
	非線形	95.0	74.0	91.6	98.6	91.6
両方無し	線形	84.7	83.9	97.6	97.7	95.7
	非線形	90.4	84.1	97.6	98.6	96.6

表 1: 実験結果のまとめ (単位 [%])　S：西高東低

なし,「西高東低」であるか否かの 2 クラス分類を行う. SVM の計算には LIBSVM[1] を用いた.

3.2 中間例

本研究で用いる正例の中には,「西高東低型」とそれ以外の型から形成される移行型・複合型が多数ある. これらの正例を中間例と定義する. 学習や検証における中間例の有無によって生じる分類精度の変化について分析する.

4 評価

4.1 実験手法

1981〜1990 年のデータを SVM の教師データとし,1991〜2000 年のデータをテストデータとした. 教師データとテストデータについて, 中間例が有無である場合の計 4 通りについて実験を行った.

4.2 実験結果

実験結果を表 1 に示す.

precision, recall はそれぞれ適合率と再現率を表す.S は「西高東低冬型」を表し, ¬ S は「西高東低冬型」を除く全てのデータを表す. 正解率は全体における正解の割合を表す.

「両方有」の場合には,precision(¬ S) と recall(¬ S) で高い数値が得られたが,recall(S) ではやや低い数値が得られた.

「学習のみ有」の場合には precision(¬ S), recall(¬ S) で高い数値が得られたが, precision(S) で低い数値が得られた.

「検証のみ有」の場合には recall(¬ S) で高い数値が得られたが,recall(S) で低い数値が得られた.

「両方無し」の場合には precision(¬ S) と recall(¬ S) で高い数値が得られたが,recall(S) が「学習のみ有」に劣り,precision(S) は「検証のみ有」に劣った.

以上の結果より, 現状では全ての評価基準について優れる手法がないことがわかった.

5 議論

実験結果からは,SVM を用いても precision(S) および recall(S) を最大でそれぞれ 95.0%, 91.7%までしか上げられないことがわかった. これは SVM の限界であるが, 人間が分類したとしても, どこまでそれらの評価値を高めることができるかわからない. なぜなら本研究では文献 [2] を用いて正例と負例を設定したが, 同書の分類手法があらゆる気象研究者に認められるとは限らないからである.

西高東低といっても非常に幅広い種類があり, 高気圧が東側に存在するパタンや, 低気圧が複数存在するパタンなど, 一見すると西高東低と判定すべきかわからないパタンが多数存在していた.

そのため, 研究者によって正例の設定は変わると考えられる. そこで気象研究者に確実に貢献するため, 見逃し (False Negative) を零にできるような手法を求めることが必要であろうと考えられる. 当然ながらその場合には False Positive は大量に発生することが予想されるが, これまで数十年間のデータを目視せざるを得なかった気象研究者の苦労を確実に減らせるだろう.

6 結論

データマイニングによる「西高東低」分類手法について述べた. SVM を用いた分類と精度の検証を行い, 最大で 95.0%だった. また, 中間例の有無によって適合率と再現率が大きく変化することがわかった.

謝辞

本研究の一部は, 科学研究費補助金萌芽研究 (# 18650018) による.

参考文献

[1] *LIBSVM.* http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[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.