第57回 プログラミング･シンポジウム 2016.1.8-10

日本の東西分割を通じた機械学習手法の評価

宮野 祐輔[1,a)] 崔 誠云[1,b)] 疋田 敏朗[1,c)] 小林 良輔[1,d)] 鈴木 宏哉[1,e)]
山口 利恵[1,f)]

概要：近年，大量のデータから機械学習によってパターンや傾向を算出するアプローチが盛んになっている．しかし機械学習手法は様々な種類が存在するため，各手法の特徴を理解しなければデータや目的に応じた適切な手法を選択することができない．そのため本研究では実際の統計データを用いた機械学習による分析を通じて，各機械学習手法の特徴を明らかにした．実験では日本人であれば事前知識があり，出力結果が妥当かどうか判断しやすい都道府県の統計データを使用した．この統計データを用いて，東日本と西日本の境界線を機械学習による判別分析によって決定した．その過程で，決定木は複雑な条件による判別は向かない，線形判別分析で得られる判別得点には有益な用途があるなど，各手法の特徴を具体的に示すことができた．このことから，都道府県の統計データのように身近で分かりやすいデータセットを用いることは機械学習手法の特徴を評価するために有用であることが分かった．

キーワード：機械学習，判別分析，線形判別分析，SVM，Random Forest

1. はじめに

近年官公庁や民間企業などにおいて，蓄積された大量のデータを利活用したサービスの新規開発や品質向上を図る動きが活発になっている．膨大で複雑なデータを処理するアプローチの一つとして，データのパターンや傾向からモデルを算出することのできる機械学習が採用されることが多い．機械学習には様々な手法が存在し，それぞれが持つ特性も大きく異なるため，同じデータを分析する場合であってもその結果や性能には差が生じることがある．したがってデータを分析する場合には，各機械学習手法の特性を十分に評価・理解した上で適切な手法を選択しなければならない．しかし実際に機械学習の対象となるデータセットは複雑で難解なものが多く，機械学習自体を評価するにあたっては適切であると言いがたい．

そこで本研究では，身近で平易な日本の都道府県別統計データを分析することを通じて，代表的な機械学習手法が持つ特性を評価した．具体的には日本の都道府県を，東日本と西日本の 2 クラス判別問題としてそれぞれ分類し，東西の境界線を導出した．今回の実験では総務省統計局が公開している人口や面積，漁獲量や農産物の収穫量，各家庭の支出金額など 246 項目の統計データを利用した．基準によって東西どちらにも分類されうる中部地方の 9 県（富山，石川，福井，山梨，長野，岐阜，静岡，愛知，三重）をテストデータ，それ以外の 38 都道府県をトレーニングデータとして，機

[1] 東京大学大学院 情報理工学系研究科
[a)] gamma@yamagula.ic.i.u-tokyo.ac.jp
[b)] song@yamagula.ic.i.u-tokyo.ac.jp
[c)] toshi@yamagula.ic.i.u-tokyo.ac.jp
[d)] kobayashi.ryousuke@sict.i.u-tokyo.ac.jp
[e)] susuki.hiroya@sict.i.u-tokyo.ac.jp
[f)] yamaguchi.rie@i.u-tokyo.ac.jp

械学習による教師あり 2 クラス判別分析を行った．実験ではテストデータの判別結果，leave-one-out cross-validation による交差検証，計算時間等の項目を求め，それらの結果を基に各手法を評価した．その後，東西格差が特に大きいと思われる 8 項目の統計データを選別して再度実験し，その結果を 246 項目で実験した場合と比較した．交差検証の結果では SVM や k-近傍法などが高い精度を示した．一方で精度の面では劣る線形判別分析でも，その結果得られる判別得点を利用し，各都道府県の「東日本らしさ」「西日本らしさ」を視覚的に表現することができた．

以上のように，本論文では複数の機械学習手法を様々な項目について評価することで各手法の長所や短所などの特性を明らかにした．また今回は身近で平易な統計データを実験に用いたことで，各手法の特性を具体的に把握することができた．

1.1 本論文の構成

本論文の構成を以下に示す．まず 2 章では，本論文中で用いた機械学習手法について説明する．次に 3 章では，東西判別を選択した理由やその特性について論じる．実験の内容と結果は 4 章，それを踏まえた考察は 5 章で説明する．最後に 6 章で本論文の結論を述べる．

2. 機械学習手法

本研究では判別問題によく用いられる代表的な機械学習手法を使って実験した．本章では各手法の判別方法や特徴について説明する．

2.1 線形判別分析

線形判別分析（Linear Discriminant Analysis; LDA）とは，式 1 に示すような線形の判別関数を学習によって作成し，判別分析を行う手法のことである．

$$\begin{aligned} f(\boldsymbol{x}) &= a_0 + a_1x_1 + a_1x_2 + \cdots + a_dx_d \\ &= a_0 + \sum_{i=1}^{d} a_ix_i \end{aligned} \tag{1}$$

このとき， $\boldsymbol{x} = (x_1, x_2, \cdots, x_d)^T (^\exists d \in \mathbb{N})$ はユーザの d 次元特徴ベクトルを表す．判別関数にユーザの特徴ベクトルを代入して得られる値を判別得点と呼び，この値を基準として判別する．特徴ベクトルの成分にはそれぞれ重み係数 $a_k (k = 1, 2, \cdots, d)$ が学習の結果から与えられる．a_0 は判別の境界を 0 に調整するための値であり，学習に用いた特徴ベクトルおよび重み係数から計算することができる．

線形判別分析は判別関数や判別得点など，シンプルで分かりやすいという長所がある．一方で多量の特徴量を持つデータを苦手とする，特徴量間の相関関係を利用できない，等分散な正規分布を仮定している，そして判別精度があまり高くないという欠点があるため実際に用いられる機会は少ない．

2.2 k-近傍法 [3]

k-近傍法（k-Nearest Neighbor; kNN）とは，テストデータとの類似度が高い k 個のトレーニングデータを選別し，クラスについて多数決をすることでテストデータのクラスを決定する機械学習手法である．類似度の指標にはユークリッド距離が用いられることが多い．

k の値は事前に決定しなければならないが，小さすぎるとノイズに対する耐性が低下し，逆に大きすぎると他クラスのデータまで含むため精度が低下してしまう．なおかつ k の値は対象となるデータによって異なるため，データごとに適切な値を設定する必要がある．k の値はヒューリスティックに決められることが多いが，交差検証を利用して最適な k を選択する手法がよく用いられている．特に $k = 1$ の場合は最近傍法と呼ばれる．

k-近傍法は与えられたテストデータに対して隣接したデータを元に判別するため，事前に学習する必要がない．その反面，学習によるモデルは形成されない．また，不必要な特徴量が多く与えられている場合，計算コストが大きくなり精度も低下することが知られている．

2.3 決定木

決定木（Decision Tree）とは，ある特徴量の値に基づきデータセットを分岐させ，それを分岐後

のサブセットに対しても再帰的に適応していく手法である．分岐構造が二分木で表わされるため，決定木や樹木モデルなどと呼ばれる．長所としては，プロセスや判別基準が明確かつ初学者にも分かりやすい点が挙げられる．一方でデータセットに僅かでも変化があると大きく木の構造が変化してしまったり，複雑な分岐条件を再現できなかったりなどのデメリットも存在する．また，後述するアンサンブル学習における弱学習器として利用されることも多い．

2.4 サポートベクターマシン [14]

サポートベクターマシン（Support Vector Machine; SVM）とは，境界面およびトレーニングデータ（2 クラス）との距離を評価関数として，マージンを最大化する超平面を判別の境界面として学習する手法である．言い換えれば，2 クラスのちょうど中間を通るような境界面を設定することで，高い汎化能力を実現している手法である．サポートベクターマシンはその優れた汎化能力と，高次元のデータに対しても高い精度を出すことから，幅広い分野で利用されている．しかしデータを高次元にするにしたがって，計算量が非常に大きくなるので留意が必要である．また，学習に使用するカーネルを選択しなければならない．

なお，SVM を利用した分類器を SVC (Support Vector Classifier) という．

2.5 アンサンブル学習

アンサンブル学習（Ensemble Learning）とは，複数の弱学習器を組み合わせることで判別分析の精度を向上させた機械学習手法である．アンサンブル学習の代表例としては，Bagging[1] や Boosting[4], [5], [6], [7], [8], [12]， Random Forest[2] などが挙げられる．これら 3 種類の手法については，詳細を以下にまとめる．

2.5.1 Bagging[1]

Bagging では，トレーニングデータをブートストラップと呼ばれる手法でサンプリングし，複数個のトレーニングデータ標本を作成する．その後それぞれのトレーニングデータ標本に対して弱学習器を用いた判別分析を行い，得られた判別結果の多数決をもって全体の判別結果とする．

Bagging はトレーニングデータをサンプリングして学習するため，並列化が可能であり高速化が望める．また弱学習器が事例を重み付けできなくても適応可能である．しかし Bagging はアンサンブル学習の中では，特に判別分析において，精度がそれほど高くないことが知られている．

2.5.2 Boosting[7], [12]

Boosting では，トレーニングデータを弱学習器によって逐次的に学習していく．そうして得られた判別精度を元に重みを更新し，新たに弱学習器を作成したうえで再学習するというサイクルを繰り返していくことで，最終的な判別結果の精度を高めていく． Boosting の代表的な手法には Adaboost[6] が挙げられる．

Boosting はトレーニングデータをすべて学習に利用できるため，データ数が少ない場合でも高い性能を発揮しやすい．一方で一定回数のサイクルを経て学習していくため並列化による高速化ができない，外れ値やミスラベルデータに対して過剰な重みを設定して過学習しやすいなどの欠点もある．

2.5.3 Random Forest[2]

Random Forest では，トレーニングデータから得られたブートストラップサンプルを元にそれぞれ決定木を作成する．それらの決定木から得られる結果を統合して最終的な判別結果を出力する．Bagging に似た手法ではあるが， Bagging がすべての説明変数を使用するのに対し， Random Forest では説明変数もランダムサンプリングするという違いがある．

Random Forest は精度やメモリ効率などの点で Bagging や Boosting より優れている．大規模データやノイズの多いデータに対しても，パラメータチューニングや事前処理を行わなくとも高い精度を示すため近年注目を集めている．弱学習器の数を増やして性能を向上させるためにはメモリを大量に使用する必要があり，極めて高い精度を求める分析には向かないことがある．また弱学習器の数を増やしすぎると過学習する可能性もあり，拡張するためには学習構造や学習方法を工夫する必

要がある．

3. 東西判別分析への適応

本章では，機械学習手法を評価する際に使用する東西日本の判別分析について述べる．

3.1 東西判別分析を選択した理由

本論文では機械学習手法の性能を評価するにあたって，実際のデータを用いて実験することとした．機械学習手法の動作確認や評価にあたっては，アヤメの計測データからなる Iris Data Set[10] や手書きの数字画像データからなる MNIST handwritten digit database[9] が良く用いられている．今回はこれらの既存データセットではなく都道府県の東西判別を選択したが，その理由を以下に記す．

- 身近で直感的に理解しやすい
  機械学習手法を評価するにあたっては，それによって得られた判別結果がどのような特性を保持しているか理解する必要がある．例えば Iris Data Set であれば，その結果の特性を十分に把握するにはデータセットに含まれる 3 種類のアヤメのサイズについての知識や経験が必要である．
  一方，日本における県民性は国内で生活していれば普段の生活の中で経験的に理解することができる．新たなデータセットを扱うためのハードルを下げるためにも，身近なデータから得られるデータセットは有用である．
- データが入手しやすい
  データセットを構築するにあたって，具体的な数値データや学術調査に基づく資料が必要である．日本の都道府県に関するデータであれば，官公庁が調査･公開しているデータや学術調査のデータなどが使用できるため，データセットを容易に構築することができる．
- 拡張が容易である
  今回は都道府県ごとの統計データからデータセットを構築したが，データが入手できるのであれば都道府県単位である必要はない．市区町村単位，旧藩単位，国家単位など，目的とする実験対象にあわせて柔軟に拡張ができるため，幅広い用途への展望がある．

3.2 既存の境界基準

本節では既存の境界線基準について整理する．

東西日本の境界線は，以前より様々な観点に基づき多くの基準が提唱されてきた．代表的な基準と分類結果を表にまとめると，表 1 の通りである．分類例のラベルは，東日本と西日本にそれぞれいくつの境界県が分類されるかを表している（東 3-6 西であれば，3 県が東日本，6 県が西日本に分類されることを示す）．

表 1 では既存の基準を**人為的基準**と**非人為的基準**の 2 種類に大別した．その土地の地理･文化･歴史などを元に意思を持って人為的に決定された基準を人為的基準と定義する．一方，調査や検証の結果として作為無く導かれる基準を非人為的基準と定義する．人為的基準を大別すると，製品や産業のローカライズのために設定されるものと，行政区分や管轄区分などのようにある組織の影響力が及ぶ範囲を明確化したものとに分けられる．前者の例としては即席麺やポテトチップスなど食品の塩分濃度が代表的であり，後者の例としては気象庁や NTT などが挙げられる．

一方非人為的基準の具体例としては，フォッサマグナや親不知などの地理的なものや，食習慣や生活様式の違いに基づく文化的なものがある．地理的な基準は時代の変化によらず具体的な形を伴って存在するため，境界線の設定が文化的な基準に比べると容易である．文化的な基準は調べる対象や調査形式，年代など様々な要素によって結果が大きく変動し，明確な正解を導くことは難しい．具体例として「東日本は鮭，西日本は鰤」という基準を考えたとき，境界候補付近に近づくにつれて同一市町村内でも大きくばらつくことが想定され

*1 新潟，富山，長野，静岡では一部 50 Hz と 60 Hz の混在地域が存在する．

*2 構造線上にある県の分類は県庁所在地に基いて決定した．静岡県は静岡市を縦断しているため，フォッサマグナとの位置関係も考慮して西日本とした．

*3 セブン-イレブンのおでんだしに関する旧基準に依った（『日本経済新聞電子版』2010 年 10 月 22 日参照）．2015 年現在ではセブン-イレブンは 8 地域でそれぞれ異なるだしを使用している，

表 1 既存の東西日本の境界線基準

分類例	山梨	長野	静岡	岐阜	愛知	富山	石川	福井	三重	人為的区分	非人為的区分
東 0-9 西	西	西	西	西	西	西	西	西	西	気象庁 [17], ポテトチップス [15]	
東 1-8 西	東	西	西	西	西	西	西	西	西	50/60Hz*1	
東 2-7 西	東	東	西	西	西	西	西	西	西	NTT[11]	糸魚川静岡構造線 *2
東 3-6 西	東	東	東	西	西	西	西	西	西	地方航空局 [19]	方言（文法）[16]
東 5-4 西	東	東	東	東	東	西	西	西	西		方言（アクセント）[16]
東 6-3 西	東	東	東	東	東	西	西	西	東	おでん *3，即席麺 [21]	
東 7-2 西	東	東	東	東	東	東	東	西	西	関ヶ原の戦い	
東 8-1 西	東	東	東	東	東	東	東	東	西	『日本国語大辞典』	
	東	東	東	東	東	東	東	西	東		雑煮の餅 [18]

る．また時代の変遷にともなって境界線が移動したり，流通の発展や食文化の変化によって鮭と鰤の地域性そのものが消失したりする可能性もある．

3.2.1 関連研究

高橋ら [23] は「鮭と鰤」「豚肉と牛肉」の消費量を利用して，東日本と西日本の判別分析を行っている．学習には各年代の鮭，鰤，豚肉，牛肉の消費量を用いて，線形判別による分析を実施している．東西日本の境界線は，新潟・長野・山梨・静岡以東を東日本，それ以外を西日本であると事前に設定している．

高橋らは事前知識を元に 2 種類の基準をあらかじめ設定し，基準にそぐわない東京や沖縄などの生活様式について考察している．本研究では複数のデータから事前知識なしに，東西日本の境界線およびその基準となる指標を求めることを目標とした．高橋らの研究が東日本・西日本という尺度から外れる都道府県やその要因をフォーカスしているのに対し，本研究ではそういった既存の手法では評価しづらい都道府県でも判別できる根拠を探索しているという点が異なっている．

4. 実験

本章では，実際のデータを用いて実施した機械学習による東西判別の実験について述べる．

4.1 実験環境

本実験では Python，および Python 上で動作する機械学習ライブラリ scikit-learn[13] を利用して実験した．動作環境は表 2 に示した．

表 2 実装・動作環境

動作言語	Python 2.7.10
ライブラリ	scikit-learn 0.16.1
OS	OS X El Capitan 10.11.1
プロセッサ	1.6 GHz Intel Core i5
メモリ	4GB 1600 MHz DDR3

本実験では政府統計の総合窓口（e-Stat）[20] で公開されている統計データのうち，都道府県別の数値が公表されているデータを利用した．そのうち国勢調査や産業別のセンサス，家計調査などから 246 項目の統計データを抽出して各都道府県の特徴ベクトルとした．

4.2 実験構成

3.2 節にて挙げた基準において，東西どちらにも分類されうる中部地方の 9 県（富山県，石川県，福井県，山梨県，長野県，岐阜県，静岡県，愛知県，三重県）を**境界県**と定義し，それ以外の 38 都道府県を**非境界都道府県**と定義した．本実験では非境界都道府県のデータをトレーニングデータ，境界県のデータをテストデータとして扱った．

まず各機械学習手法に基づく学習器を，デフォルトのパラメータを用いて作成した．その後，非境界都道府県のトレーニングデータ 246 項目，および東日本・西日本のラベル 1 項目を要素に持つベクトルを学習器に入力し，学習させた，そうして得られた学習済み学習器に対して，境界県のテストデータ 246 項目から東日本・西日本のラベルを推測させた．この結果を記録し，機械学習手法ごとに集計した．

また，統計データ 246 項目の中から特に東西によって差の出やすい 8 項目（牛・豚・鶏，鮭・鰤・鯛，納豆，グレープフルーツの各消費量）を選別して標準化した後，同様の実験を実施した．

以上の結果を複数の評価基準と照らしあわせ，各機械学習手法の長所・短所，特徴などを分析した．

4.3 予備実験

k-近傍法の k の値は，データセットのサイズや性質によって適切な値が大きく異なってくる．そのため k-近傍法を利用する際の k は，トレーニングデータを使った交差検証によって適切な値をヒューリスティックに求めて使用することが多い．したがって今回は k を 1 から 37 まで変化させた上で LOOCV を実施し，最も誤判別の少ない k を実験に使用することとした．

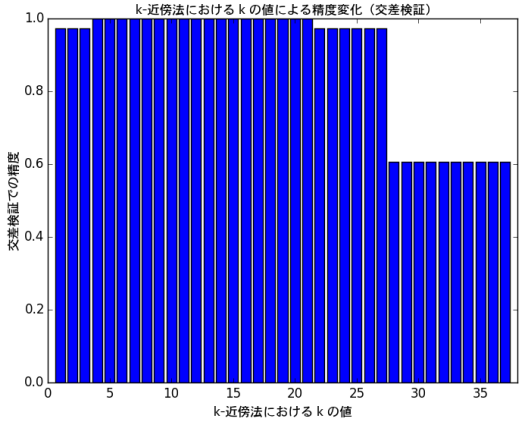

図 1 k-近傍法における k の値による精度変化（交差検証）

予備実験の結果を図 1 に示す．$k \geq 28$ において精度が急激に低下しているのは，k が大きすぎるためにトレーニングデータがすべて多数派である西日本と判別されたためである．今回はトレーニングデータ数が 38 と少なかったことから k が幅広い範囲（$4 \leq k \leq 21$）で誤判別が発生せず，適切な k の値をこの結果からだけでは判断することは難しい．そのため今回は scikit-learn のデフォルト値 である $k = 5$ で誤判別がなかったことから，他の機械学習手法同様デフォルト値である $k = 5$ を選択して学習器を作成した．

4.4 実験結果

4.4.1 各手法による境界県の判別結果

各手法による境界県の判別結果を表 3，4 に示した．あわせて非境界都道府県に対して実施した LOOCV での誤判別率と，誤った都道府県の数も記載した．誤判別した都道府県が 1 つだけであれば，数ではなく都道府県名を記載した．なお，SVC は 246 項目を使用した場合すべて西日本と判別するような学習をしてしまった．そのため SVC については 8 項目を使用した場合についてのみ述べる．

4.4.2 決定木

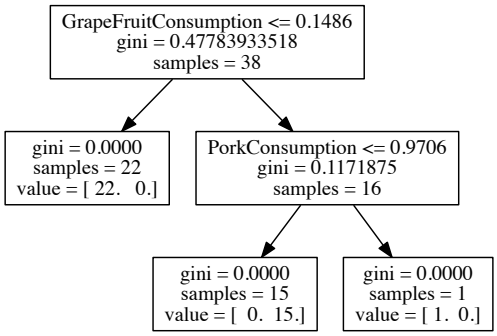

図 2 決定木（$d = 8$）

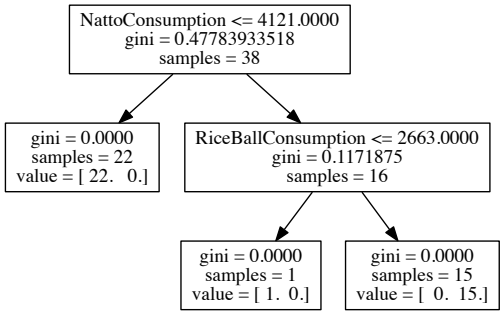

図 3 決定木（$d = 246$）

決定木判別から得られた結果を元に，図 2，3 の二分木を描画した．二分木のノードには，その

表 3 機械学習を用いた境界県の判別結果（8 次元）

機械学習手法	山梨	長野	静岡	岐阜	愛知	富山	石川	福井	三重	LOOCV ($n = 38$) での誤判別率
LDA	東	東	東	西	西	東	西	西	西	2.63%（沖縄）
k-近傍法	東	東	東	西	西	西	西	西	西	0.00%
決定木	東	東	西	西	西	東	東	西	西	13.16%（5 都県）
SVC	東	東	東	西	西	西	西	西	西	2.63%（沖縄）
Bagging	東	東	西	西	西	東	西	西	西	2.63%（東京）
Adaboost	東	東	西	西	西	東	西	東	西	7.89%（3 都県）
Random Forest	東	東	西	西	西	東	西	西	西	2.63%（東京）

表 4 機械学習を用いた境界県の判別結果（246 次元）

機械学習手法	山梨	長野	静岡	岐阜	愛知	富山	石川	福井	三重	LOOCV ($n = 38$) での誤判別率
LDA	西	東	西	西	東	東	西	東	西	0.00%
k-近傍法	西	東	西	東	東	西	西	西	西	34.21%（13 府県）
決定木	東	東	西	西	西	東	西	東	西	7.89%（3 都県）
Bagging	東	東	西	西	西	東	西	西	西	28.95%（11 都府県）
Adaboost	東	東	西	西	西	東	西	東	西	10.53%（4 都県）
Random Forest	西	東	西	西	西	西	西	西	西	13.16%（5 都県）

ノードに含まれるサンプル数と，決定木の評価値であるジニ係数を記載した．末端のノードにはそのノードに属するサンプルが持つ真のラベルの割合が，それ以外のノードには分岐条件をそれぞれ記載した．

4.4.3 LDA の判別得点を利用した東日本度・西日本度の算出

LDA の特徴として，他の学習器のように単純なクラス分類だけではなく，その際に算出された判別得点が得られるというものがある．今回の実験では判別得点が正であれば東日本，負であれば西日本と判別した．したがって判別得点の絶対値が大きければその都道府県の東日本・西日本らしさが強く，絶対値が小さければ東日本・西日本らしさが弱い都道府県であるといえる．

本節では東日本･西日本らしさを表す指標を設定し，東日本･西日本度（$Eastness_n, Westness_n$）とする．このとき n は JIS X0401[22] に定められた各都道府県の都道府県番号である（$n = 1, 2, \cdots, 47$）．各都道府県の判別得点を f_n とし，それらのうち最大値を f_{max}，最小値を f_{min} としたとき，都道府県番号が n である都道府県の東日本度・西日本度は以下のように表される．

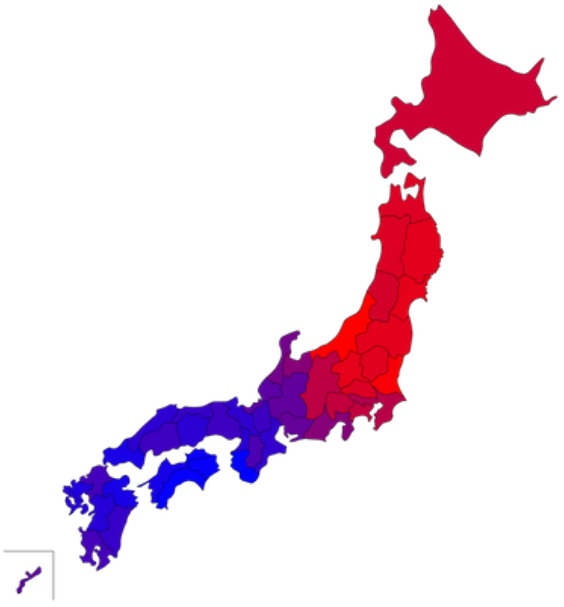

図 4 算出した東日本度・西日本度に基づく日本地図

$$Eastness_n = \frac{f_n - f_{min}}{f_{max} - f_{min}} \quad (2)$$

$$Westness_n = 1 - Eastness_n \quad (3)$$

各都道府県の色を東日本度が大きいほど赤く，西日本度が大きいほど青くなるように色分けしたのが図 4 である．

4.4.4 Random Forest における各特徴量の寄与度

Random Forest では，判別に利用した各特徴量の寄与度が算出可能である．寄与度が大きいほど判別にとって重要な特徴量であったということができ，あるいは寄与度が 0 である特徴量は，Random Forest にとっては 判別に利用しない不必要な特徴量であるということを示す．

246 項目を利用した学習の結果，22 項目に 0 でない寄与度が与えられた．表 5 にはそのうち寄与度の大きい 10 項目について，最大値を 100%として算出した値を記載した．

表 5 Random Forest における各特徴量の寄与度（246 項目中，上位 10 項目）

特徴量	寄与度（%）
サンマの消費量	100
グレープフルーツの消費量	83.2
鮭の消費量	83.2
お菓子の消費量	82.6
イワシの消費量	58.9
スイートコーンの生産量	58.6
年間晴れ日数	53.4
男女性比	48.8
果物の消費量	47.9
豚肉の消費量	46.4

4.4.5 実行時間

各手法を用いたプログラムを 100 回実行し，その所要時間を箱ひげ図を用いて図 5，6 に示した．アンサンブル学習は他の機械学習手法に比べて実行時間が長かったため，他の手法と分離して描画した．グラフの横軸は使用した機械学習手法であり，末尾の数字が使用したデータセットの次元数を表している．縦軸は実行時間であり，単位は msec を使用した．

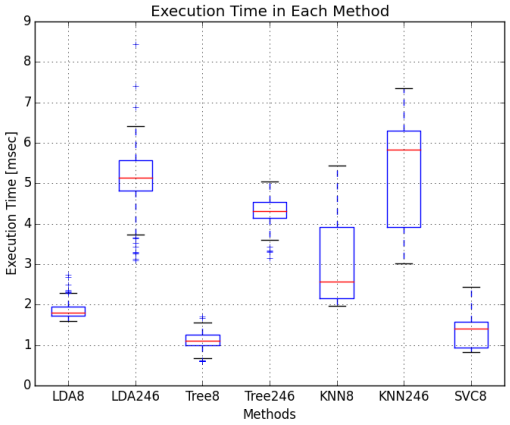

図 5 各手法プログラムの実行時間

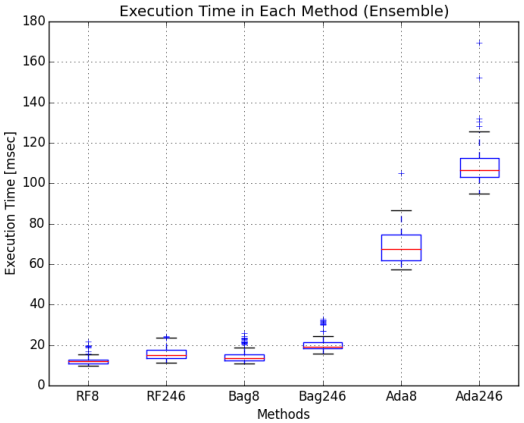

図 6 各手法プログラムの実行時間（アンサンブル学習）

5. 評価および考察

本章では，4 章で実施した実験の結果を評価し，その値および分析手法について考察する．

5.1 評価時間についての考察

アンサンブル学習は複数の弱学習器を組み合わせて判別するため，一般的に他の手法より実行時間が長くなった．特に Adaboost は逐次的に学習するため高速化が困難であり，その結果他の手法と比較して十数倍〜数十倍実行時間が長くなっている．アンサンブル学習以外の手法では，特に k-近傍法が実行時間が長い．このことは（特に特徴量が多い場合に）計算コストが大きいという k-近傍法の特徴を表したものだといえる．

5.2 評価基準

前述してきた通り，東西日本の境界線に明確な基準や正解は存在しない．しかし一定の基準を設けずに結果の妥当性を評価することは困難である．そこで評価に用いる以下の基準を設定した．

(1) 交差検証における精度

非境界都道府県に対して leave-one-out cross-validation (LOOCV) を実施し，その精度を元に機械学習による予測・分類の妥当性を評価した．具体的には非境界都道府県の中から 1 都道府県を抽出し，それ以外の 37 都道府県をトレーニングデータとして学習する．学習結果をもとに抽出した 1 都道府県の東西判別を行い，それを他の都道府県に対しても繰り返していく．38 回の検証の結果，誤識別が少なければ少ないほど学習に用いた手法はデータから高い精度の学習ができているといえる．

(2) 飛び地の有無

境界県の判別を各手法に基づき決定したとき，飛び地が存在するかどうかで境界県の判別予想結果の妥当性を評価した．飛び地が存在する場合，「東西日本を大きく二分する境界線」という概念にそぐわないと考えられるため，妥当性を低く評価した．

(3) 既存基準との比較

多く存在する既存の境界線の中から，事前に最も妥当であると考えられる境界線を選択して評価基準とした．本研究では 3.2 節および表 1 で述べた基準のうち「**長野，山梨，静岡以東が東日本（富山，岐阜，愛知以西が西日本）**」を評価基準とした．その理由は以下に示す通りである．

- 東 4-5 西という基準が得られていないこと
  表 1 において，4 つの境界県が東日本に分類される東 4-5 西という基準が得られていない．この原因として，岐阜と愛知が文化的な繋がりが深く，東西で分断されることがないという点が挙げられる．
- 東 3-6 西の前後の基準が曖昧であること
  表 1 において，東 3-6 西を採用しているのは，地方航空局の管轄と方言（文法）の 2 点である．しかし東 2-7 西の基準となっている NTT の管轄 および 糸魚川静岡構造線に関しては，いずれも静岡県をまたいだ形で存在しており，静岡が東に属する（東 3-6 西）か西に属する（東 2-7 西）かどうかは評価者の判断によって異なってくる．
  また，東 5-4 西の基準である方言（アクセント）も複数のアクセントから総合的に判断した場合であり，いくつかのアクセントについては岐阜や愛知を西に属させる（東 3-6 西）ような分布をしているものも存在する．これらの理由により，東 2-7 西や東 4-5 西の境界線基準も判断によっては東 3-6 西の境界線基準になる可能性を持つことが分かる．
- 隣接している基準の種別に偏りが小さいこと
  今回の検証では様々な統計データを利用して，最も汎用性の高い東西日本の境界線を設定しようということを目標としている．したがって基準となる指標の種別も多く存在することが望ましい．東 3-6 西付近では人為・非人為，地理的・行政的・文化的基準が様々密集しており，この基準を満たしている．

(4) 地域性の考慮

日本の都道府県の中には，隣県との間に強固な関係性を有しているものも存在する．3.2 節

でも述べた愛知県と岐阜県はその代表例であり，産業的にも文化的にも密接に関係している．今回は**長野・山梨（甲信），愛知・岐阜，富山・石川（旧加賀藩）**の組を強固な関係性を有していると設定し，この地域性を分断するような判別を下す学習手法を低く評価した．

5.3 判別結果についての考察

本節では，機械学習の結果得られた判別分析の結果について考察する．

5.3.1 線形判別分析

線形判別分析では，246 次元データを使用した際の誤判別率が 0%となっているが，愛知県や福井県などが飛び地になってしまった．特に愛知県と岐阜県が分断されたため，この結果の妥当性は低い．誤判別率が 0%になったのは特徴量が多くなったことで過学習したのではないかと考えられる．

5.3.2 *k*-近傍法

k-近傍法においても，8 次元データを使用した際の誤判別率が 0%となっているが，こちらには飛び地や分断がなく，かつ 5.2 節の (3) で採用した評価基準とも一致している．一方で 246 次元データを使用した場合には誤判別率の上昇や飛び地の発生など著しく性能が低下していることから，次元の高いデータに対する *k*-近傍法の弱さが裏付けられた．決定木においては誤判別率や導出された境界線など性能としては他の手法に劣るが，二分木によって判別規則を視覚的に表現することができた．

5.3.3 SVC

SVC では 246 次元データこそ使用できなかったものの，8 次元データでは誤判別率も低く 5.2 節の (3) で採用した評価基準とも一致している．

5.3.4 アンサンブル学習

Bagging はサンプリングしたデータから学習したために，高次元のデータでは十分に学習することができなかったと考えられる．今回利用した都道府県単位の東西判別ではトレーニングデータが 38 個しか用意できなかったのも要因の一つと考えられる．一方で Boosting ではトレーニングデータ全体を，Random Forest では特徴量も含めたサンプリングするために Bagging に比べてデータの次元による判別結果への影響は小さい．しかし飛び地なども多く発生しており，Random Forest（8 次元）以外の結果は妥当性が低いと考えられる．

以上を踏まえた上で，本研究における東西日本の境界線を決定した．決定に際しては精度の高い 8 次元データを利用した上で，誤識別率が低い手法の結果の中から多数決によって選抜した．その結果「山梨，長野，静岡が東日本」「山梨，長野，富山が東日本」という両案が同率で並ぶものの，(1) 平均誤識別率が低い（0%の *k*-近傍法を含む） (2) 既存の基準とも合致するということから，**「山梨，長野，静岡が東日本」**という基準を本研究における東西日本の境界線として決定した．

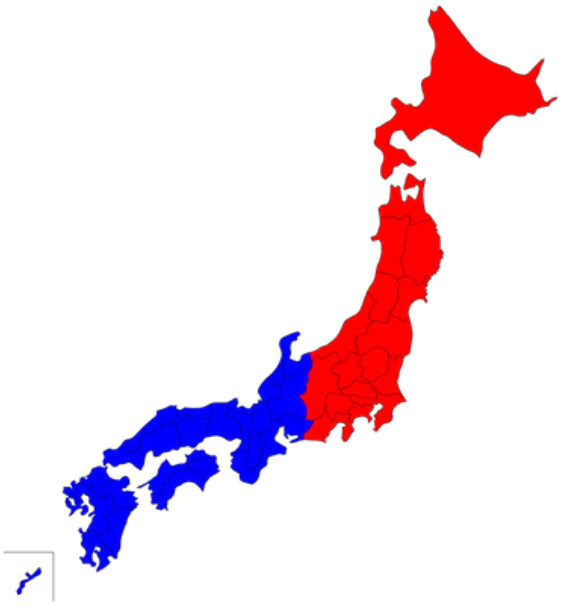

図 7 本論文における境界線基準に基づく東西日本

5.4 特徴的な都道府県についての考察

実験の結果，他の都道府県に比べて予想とは異なる結果となった都道府県が複数存在する．本節ではその原因について考察する．

各機械学習手法による東西判別の結果，特に東京都と沖縄県は LOOCV において誤判別される確率が高かった．東京都は地方からの移住者が多く流入してくることもあって，都民性が統計データに現れにくいために誤判別されやすい．沖縄県は

西日本に分類しているものの，地理的にも文化的にも本土とは乖離しているために同一の尺度で判断するのは困難である．沖縄ではラフテーやミミガーなど，「ひづめと鳴き声以外は全部食べる」と言われるほど豚肉を大量に消費することも誤判別のさらなる要因となっている．

また，富山県は既存の東西基準では西日本と判別される割合が多かったものの，機械学習手法での東西判別では東日本と判別される割合が多かった．これは富山県が呉羽山という富山市にある山を境として呉東・呉西と区分され，両者の間で文化に隔たりが存在するためであると考えられる [24]．今回使用したデータの多くは都道府県庁所在地のものを使用している．そのため県庁所在地である富山市で東西の文化が混在している状況が判別結果に影響したのではないかと考えられる．この現象は糸魚川静岡構造線が静岡市を通っている静岡県でも観測されている．

6. 結論

本論文では複数の機械学習手法を，都道府県単位の統計データから東西判別を行うことで各手法を評価した．その結果各手法の実行時間，精度，長所・短所などの特徴を調査することができた．機械学習を用いて分析する際には，データサイズ，次元数，時間制約などの観点から適切な手法を選択する必要があることが分かった．また，「山梨，長野，静岡が東日本」という既存の基準とも合致する結果が得られたことから，機械学習手法の有用性を確かめることができた．

サンプル数が少なく妥当性の評価に疑問が残ったため，市区町村単位や国家単位の統計データを利用することでより妥当な各手法の評価することは今後の課題とした．また今回採用しなかったニューラルネットワークなどの手法を追加する，別のライブラリや言語・ソフトウェアを利用してそれぞれの差異を評価するといったアプローチもさらなる理解のために有効であると考えられる．

謝辞 本研究は次世代個人認証技術講座（三菱UFJ ニコス寄附講座）の助成を受けて実施された．ここに謝意を表する．

参考文献

[1] Breiman, L.: Bagging Predictors, *Mach. Learn.*, Vol. 24, No. 2, pp. 123–140 (1996).
[2] Breiman, L.: Random Forests, *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32 (2001).
[3] Cover, T. and Hart, P.: Nearest Neighbor Pattern Classification, *IEEE Trans. Inf. Theor.*, Vol. 13, No. 1, pp. 21–27 (2006).
[4] Demiriz, A., Bennett, K. P. and Shawe-Taylor, J.: Linear Programming Boosting via Column Generation, *Mach. Learn.*, Vol. 46, No. 1-3, pp. 225–254 (2002).
[5] Freund, Y.: An Adaptive Version of the Boost by Majority Algorithm, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, pp. 102–113 (1999).
[6] Freund, Y. and Schapire, R. E.: Experiments with a New Boosting Algorithm, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, pp. 148–156 (1996).
[7] Freund, Y. and Schapire, R. E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *J. Comput. Syst. Sci.*, Vol. 55, No. 1, pp. 119–139 (1997).
[8] Friedman, J., Hastie, T. and Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting, *Annals of Statistics*, Vol. 28, p. 2000 (1998).
[9] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).
[10] Lichman, M.: UCI Machine Learning Repository (2013).
[11] NTT 東日本：プロフィール． `http://www.ntt-east.co.jp/aboutus/profile.html` 2015 年 11 月 9 日閲覧．
[12] Schapire, R. E. and Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions, *Machine Learning*, pp. 80–91 (1999).
[13] scikit-learn developers: scikit-learn: machine learning in Python. `http://scikit-learn.org/` 2015 年 11 月 9 日閲覧．
[14] Vapnik, V. N.: *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc. (1995).
[15] カルビー：ニュースリリース． `http://www.calbee.co.jp/newsrelease/070122b_2.html` 2015 年 11 月 9 日閲覧．
[16] 清哉安部：方言区画論と方言境界線と方言圏の比較研究，人文， Vol. 13, pp. 21–55 (2015).
[17] 気象庁：予報用語地域名． `http://www.jma.go.jp/jma/kishou/know/yougo_`

hp/tiikimei.html 2015 年 11 月 9 日閲覧.
[18] 畑江敬子，飯島久美子，小西史子，綾部園子，村上知子，香西みどり：正月の雑煮の食べ方に関する実態調査，日本調理科学会誌，Vol. 36, No. 3, pp. 234–242 (2003).
[19] 国土交通省：地方航空局. http://www.mlit.go.jp/about/chihokoku.html 2015 年 11 月 9 日閲覧.
[20] 総務省統計局：政府統計の総合窓口 (e-Stat). http://www.e-stat.go.jp/ 2015 年 11 月 9 日閲覧.
[21] 日清食品：こだわり｜こだわりのつゆ. http://www.donbei.jp/kodawari/index02.html 2015 年 11 月 9 日閲覧.
[22] 日本工業標準調査会：JIS X 0401 都道府県コード (1970).
[23] 高橋洋子，小谷スミ子：東日本・西日本における「鮭と鰤」「豚肉と牛肉」の購入量，一般社団法人日本家政学会研究発表要旨集，Vol. 58, pp. 4–4 (2006).
[24] 高山龍太郎：富山県の東西における地域差：富山市と高岡市のサーベイ調査から，富山大学紀要. 富大経済論集， Vol. 52, No. 3, pp. 603–652 (2007).