法数学のための機械学習の基礎

京大(医)　統計遺伝学分野

山田　亮

2017/04/15

Forensic Science International: Genetics

Volume 27, March 2017, Pages 82–91

Research paper

PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures

Michael A. Marciano[1], Jonathan D. Adelman[1]

Show more

http://dx.doi.org/10.1016/j.fsigen.2016.11.006

簡単に言うと

従来法
MAC(Minimum Allele Count)
MLE
MCMC

- 混合試料
- 構成人数推定
- 機械学習手法では、従来法に比較して
- 推定結果が良くなった
  - ３人混合の場合
  - ４人混合の場合

使われた機械学習手法

- k-NN (k-Nearest Neighbors, 近いほうからk番までを使う法)
- CART (Classification and Regression Trees、決定木法の１つ)
- Multinomial Logistic Regression
- Multilayer Perceptron (ニューラル・ネットワーク)
- SVM (Support Vector Machine) (超平面分離)

  - Python Scikit-learn ライブラリ
  - WEKA

評価手順

- たくさんの「真の答え」がわかっている混合試料とその実験結果を用意(1405試料)
- 人数推定に有用な指標を選別
  - 候補指標を挙げ、その指標が有用かどうかを「正解に照らして」選別
- 有用指標を用いて『機械学習』
  - 「正解アリ」の試料-実験結果で、判定ルールを『学習』する
  - 『学習結果』を使って、人数推定する
    - 「隠してあった答え」を見て、推定結果の良さを確認する
- 『機械学習させた分類器』に、「推定人数」の確率を出力させる(Probabilistic Assessment… “PACE”)

D_{KL}	features
1.638	sample-wide peak count
1.308	maximum number of contributors
1.060	minimum number of contributors
0.823	template DNA amplified
0.512	locus-specific peak count
0.358	min/max observed peak heights
0.309	probability of dropout
0.090	minimum observed peak height
0.038	maximum observed peak height
0	size of locus

Table 2.
Summary metrics for six machine learning algorithms' learned models for number of contributor classification. Training and testing accuracies are used to evaluate model convergence, and represent total accuracy across all 4 classes. Hyperparameter tuning was limited to hyperparameters impacting model variance, and the reported metrics describe optimized models.

Classifier	Number of Contributors	Precision	Recall	f1-score	Informedness	Training/testing Accuracy
k-NN	1	0.96	0.99	0.98	0.940	0.981/0.955
	2	0.98	0.97	0.97		
	3	0.98	0.87	0.92		
	4	0.79	0.98	0.88		
CART	1	0.97	1.00	0.98	0.965	0.974/0.975
	2	0.98	0.98	0.98		
	3	0.99	0.93	0.96		
	4	0.93	0.98	0.96		
Logistic regression	1	0.97	0.98	0.97	0.949	0.963/0.961
	2	0.97	0.98	0.98		
	3	1.00	0.89	0.94		
	4	0.83	1.00	0.90		
MLP	1	0.97	0.96	0.96	0.943	0.970/0.962
	2	0.96	0.97	0.96		
	3	0.96	0.95	0.96		
	4	0.95	1.00	0.97		
SVM (linear)	1	0.91	0.96	0.94	0.842	0.912/0.894
	2	0.89	0.90	0.89		
	3	0.89	0.77	0.82		
	4	0.88	0.96	0.92		
SVM (non-linear)	1	0.96	0.99	0.97	0.957	0.982/0.971
	2	0.98	0.97	0.97		
	3	1.00	0.93	0.96		
	4	0.93	1.00	0.96		

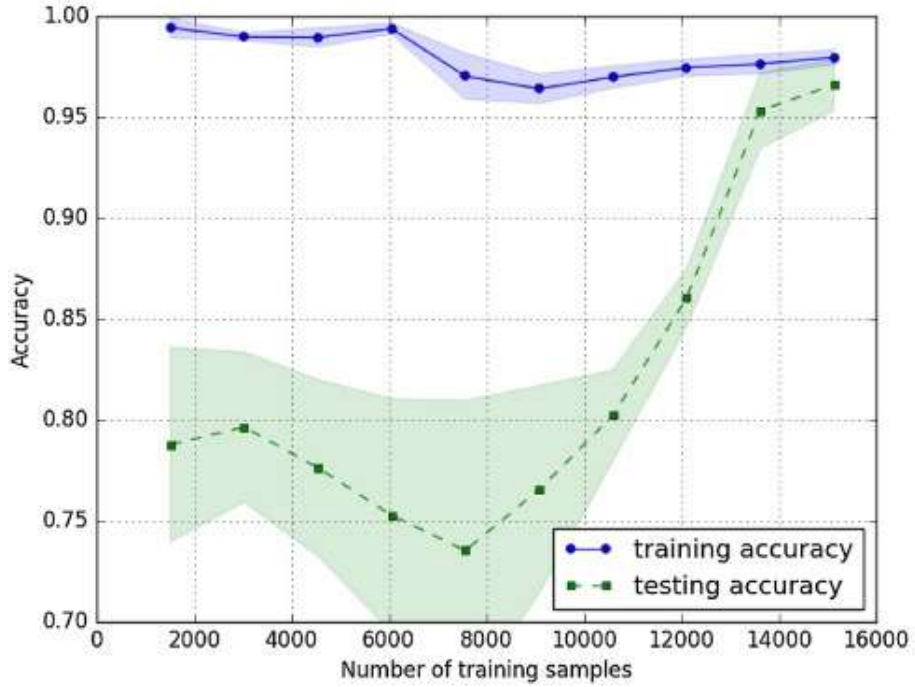

Fig. 1. Learning curve for a number of contributor estimation model derived from a support vector machine with a Gaussian kernel. The shaded area represents one standard deviation. Testing accuracy: 0.980. Note: the number of samples in this figure represent t...

Michael A. Marciano, Jonathan D. Adelman

PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures

Forensic Science International: Genetics, Volume 27, 2017, 82–91

http://dx.doi.org/10.1016/j.fsigen.2016.11.006

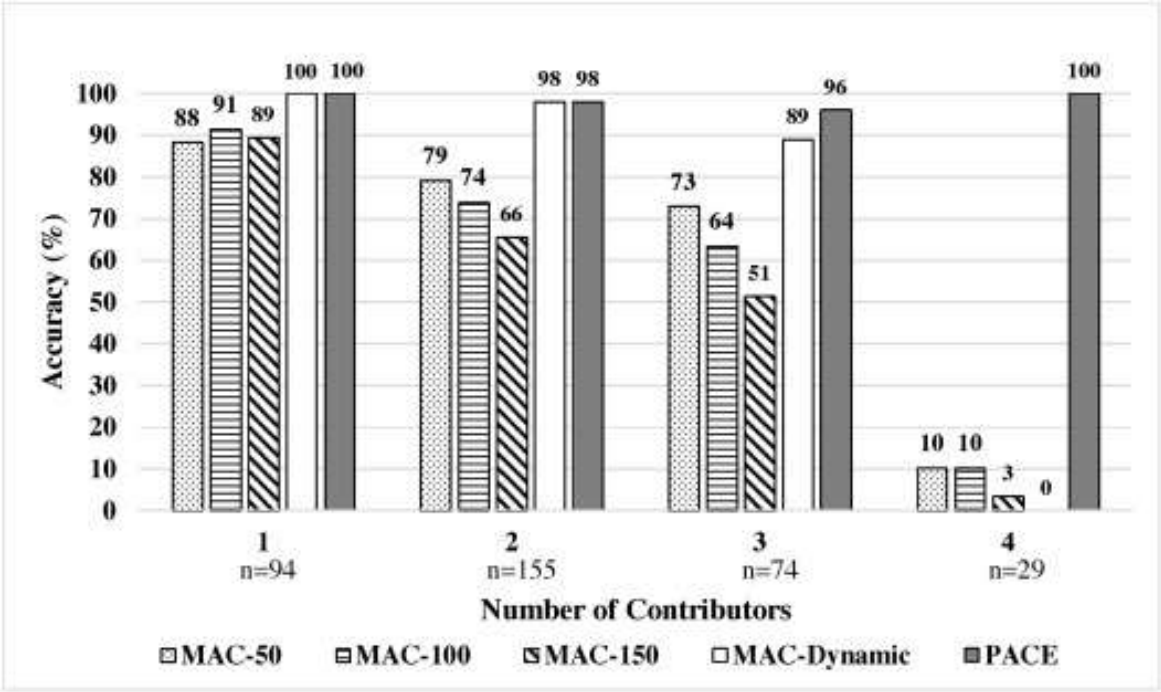

Fig. 2. Accuracy rates for several number of contributor estimation models, PACE (dynamic threshold), MAC at 50rfu, 100rfu, 150rfu and dynamic threshold.

Michael A. Marciano, Jonathan D. Adelman

PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures

Forensic Science International: Genetics, Volume 27, 2017, 82–91

http://dx.doi.org/10.1016/j.fsigen.2016.11.006

機械学習のまとめ

- 手法：いろいろある
- 教師アリ(と教師ナシ)
- 過学習(オーバーフィッティング)
- トレーニングとクロス-バリデーション

“統計学”と“機械学習”とは

- 同じだけれど、用語が違う！
- 数学寄りの統計学
- アルゴリズム好きな機械学習、情報学的な機械学習

Table 2.

Summary metrics for six machine learning algorithms' learned models for number of contributor classification. Training and testing accuracies are used to evaluate model convergence, and represent total accuracy across all 4 classes. Hyperparameter tuning was limited to hyperparameters impacting model variance, and the reported metrics describe optimized models.

Classifier	Number of Contributors	Precision	Recall	f1-score	Informedness	Training/testing Accuracy
k-NN	1	0.96	0.99	0.98	0.940	0.981/0.955
	2	0.98	0.97	0.97		
	3	0.98	0.87	0.92		
	4	0.79	0.98	0.88		
CART	1	0.97	1.00	0.98	0.965	0.974/0.975
	2	0.98	0.98	0.98		
	3	0.99	0.93	0.96		
	4	0.93	0.98	0.96		

統計系の用語と学習系の用語

		予想 (+),(-)			
	集団全体	予想(+)	予想(-)	有病率・全体に占める(+)の割合 $= \frac{\Sigma \text{真に(+)}}{\Sigma \text{集団全体}}$	
真の状態 (+),(-)	真に (+)	**真陽性 (TP)**	**偽陰性 (FN)** (第二種過誤)	真に (+)であって 予想も(+)である割合 (真陽性率 TPR), 感度, リコール, 検出する確率 $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	真に (+)であるのに 予想が(-)である割合 (偽陰性率 FNR), 見逃し確率 $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	真に (-)	**偽陽性 (FP)** (第 1 種過誤)	**真陰性 (TN)**	真には (-)であるのに 予想が(+)である割合 (偽陽性率 FPR), フォールアウト(誤って拾われる率), Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	真に (-)であって 予想も(-)である割合 (真陰性率 TNR), 特異度 Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
	正答率Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	陽性 的中 率 (PPV), プレシジョン $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	誤って 落とす 率 (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	陽性 尤度 比 (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	診断的 オッズ 比 (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		フォールス・ディスカバリ・レート 誤って 取り込んでしまう率 (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	陰性 的中 率 (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	陰性 尤度 比 (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

- 機械学習とは・・・
  https://www.slideshare.net/hayatomaki9/litmachinelearning

- 各論・・・