



属性付き中心順序の推定 手法のサーベイと比較実験

神鳶敏弘(産総研) 賀沢秀人(NTT) 赤穂昭太郎(産総研)
<http://www.kamishima.net/>
第18回 人工知能学会 全国大会 (2004.6.3)

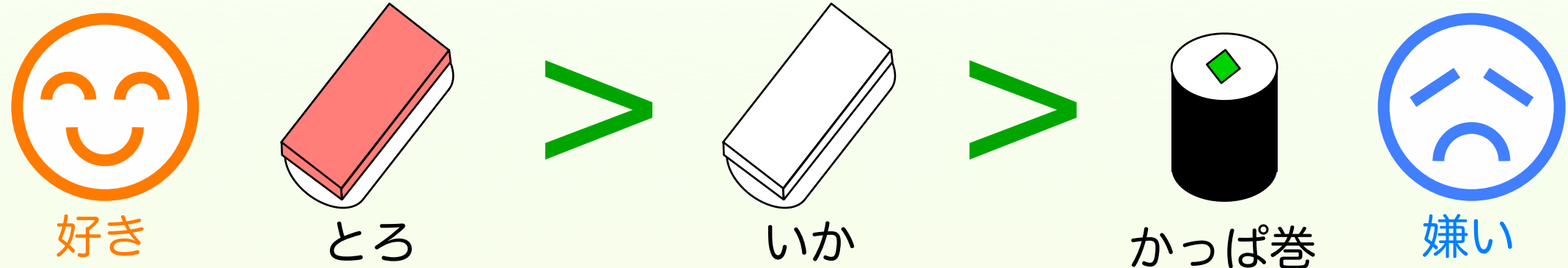


概要

訓練事例として与えられた順序とできるだけ一致する
ように対象を整列する規則の学習問題

順序：何らかの基準で対象を整列したもの

例：Aさんが好き[基準]な順に寿司[対象]を整列



“いか”より“とろ”が好き
しかし、どれくらい好きかは不明

5種類の手法について比較実験

なぜ「順序」なのか？

広い適用範囲

- ▶ 頻繁に利用されているデータの表現手法

検索結果の表示，売り上げなどのランキングなど

- ▶ 主観的な変量の計測に適す

嗜好や感覚など主観的な変量は，定量的に測るより，相対的な大小関係を測る方が容易

重要な測度の水準

- ▶ 順序は選択や決定にとって本質的な量

選択肢のスコアのエラーがいくら小さくても，選択肢の適切さの順序は保証されない

属性付き中心順序の推定

訓練用の順序

$A > B > C$
 $D > E > B > C$
 $A > D > C$

将来発生する順序

$A > B > C$
 $E > D > A > B$
 $D > C > B$

どちらにも平均的に
類似した順序

中心順序

$E > D > A > B > C$

$B \ C \ E$

未整列の対象

整列関数

対象の並びが一致

$E > B > C$

中心順序と一致した順序

訓練用の順序から整列関数を学習する問題
対象が属性ベクトルで記述されている場合が
「属性付き中心順序の推定」問題

Cohenの方法 [Cohen 99]

学習: 属性値から対象対の順序関係を推定する確率関数

訓練用の順序

$A \succ B \succ C$
 $D \succ E \succ B \succ C$
 $A \succ D \succ C$

一対比較に分解

$A \succ B, A \succ B, B \succ C$
 $D \succ E, D \succ B, D \succ C, \dots$
 $A \succ D, A \succ C, D \succ C$

確率関数の学習

$\Pr[x \succ y | x, y]$

整列: 未整列の対象を評価関数に基づき整列

$$\sum_{x \succ y \in \text{Pair}} \Pr[x \succ y | x, y]$$

評価する順序

$A \succ B \succ C$

一対比較に分解

$\text{Pair} = \{A \succ B, A \succ B, B \succ C\}$

欲張り探索で最良の順序を見つける

SVM系手法

学習: 対象が上位なる度合いを表す効用関数を獲得

▶ カーネル用いた非線形の効用関数を利用可能

▶ 次の条件を満たす効用関数をSVMを用いて計算

Order SVM (OSVM) [賀沢 03]

ある対象の順位が j 番目より上位になるかどうかを正しく判別

Support Vector 順序回帰 (SVOR) [Herbrich 98]

二つの対象のどちらが上位に配置されるかを判別

整列: 効用関数を利用

未整列の対象を, その属性値を与えた効用関数の値で整列

Thurstone系手法

学習: 対象が上位なる度合いを表す効用関数を獲得

▶ 現在は線形の効用関数のみ

▶ 「Thurstoneの比較判断の法則」のモデルを利用

対象 x の効用値は正規分布 $N(\mu_x, \sigma)$ に従い, 順序は効用値で整列することで生じるとするモデル

Order SVM (TR) [神嶌 02] 最少二乗推定

属性付きThurstoneモデル (ATM) [赤穂 02] 最尤推定

整列: 効用関数を利用

未整列の対象を, その属性値を与えた効用関数の値で整列

人工データでの実験

データの生成手順

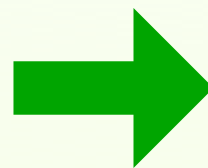
1. 対象の属性値をランダムに生成
2. 真の中心順序を次の効用関数に基づいて生成

$$\text{utility}(x_j) = (1 + \sum_l^k w_l x_{jl})^{\text{dim}}$$

3. 真の中心順序から, 順序集合をサンプリング

真の中心順序

$A \succ B \succ C \succ D \succ E \succ F$



標本順序の集合

$A \succ C \succ D \quad B \succ C \succ E \succ F$

テスト方法: 交叉確認

1. 順序集合を訓練用とテスト用に分ける
2. 訓練用集合から整列関数を学習
3. テスト用の集合で, 整列関数で予測した順序と真の順序を比較

順序ノイズと属性ノイズ

順序ノイズ

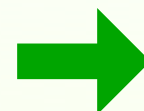
ノイズなし

真の中心順序

$A \succ B \succ C \succ D$

対象はサンプリングされても

対象間の順序関係は保存



標本順序

$A \succ B \succ D$

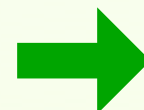
ノイズあり

真の中心順序

$A \succ B \succ C \succ D$

対象はサンプリングされて

さらに、順序の置換も加わる



標本順序

$D \succ B \succ A$

ランダムに選んだ隣接する対象を交換



属性ノイズ

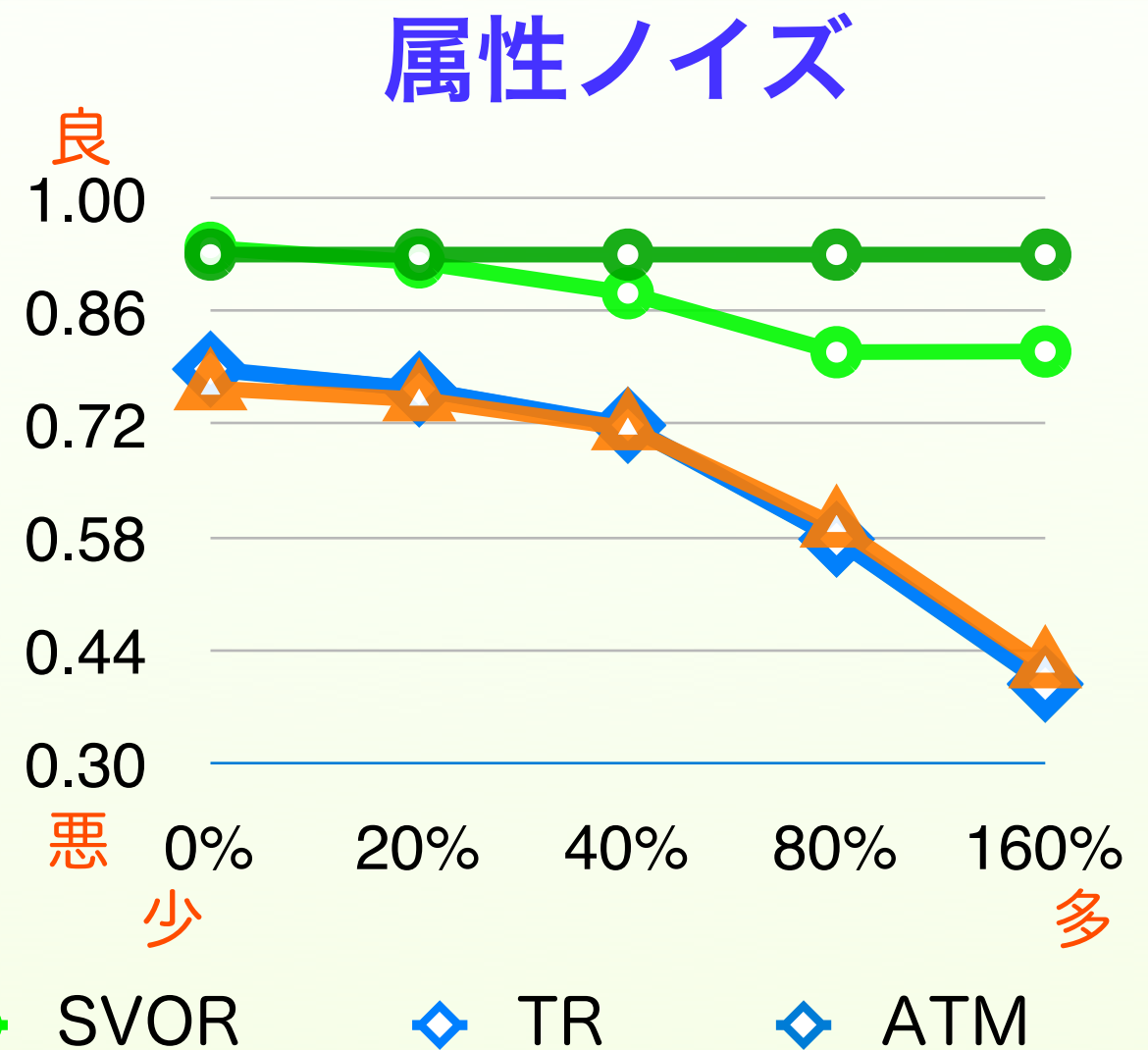
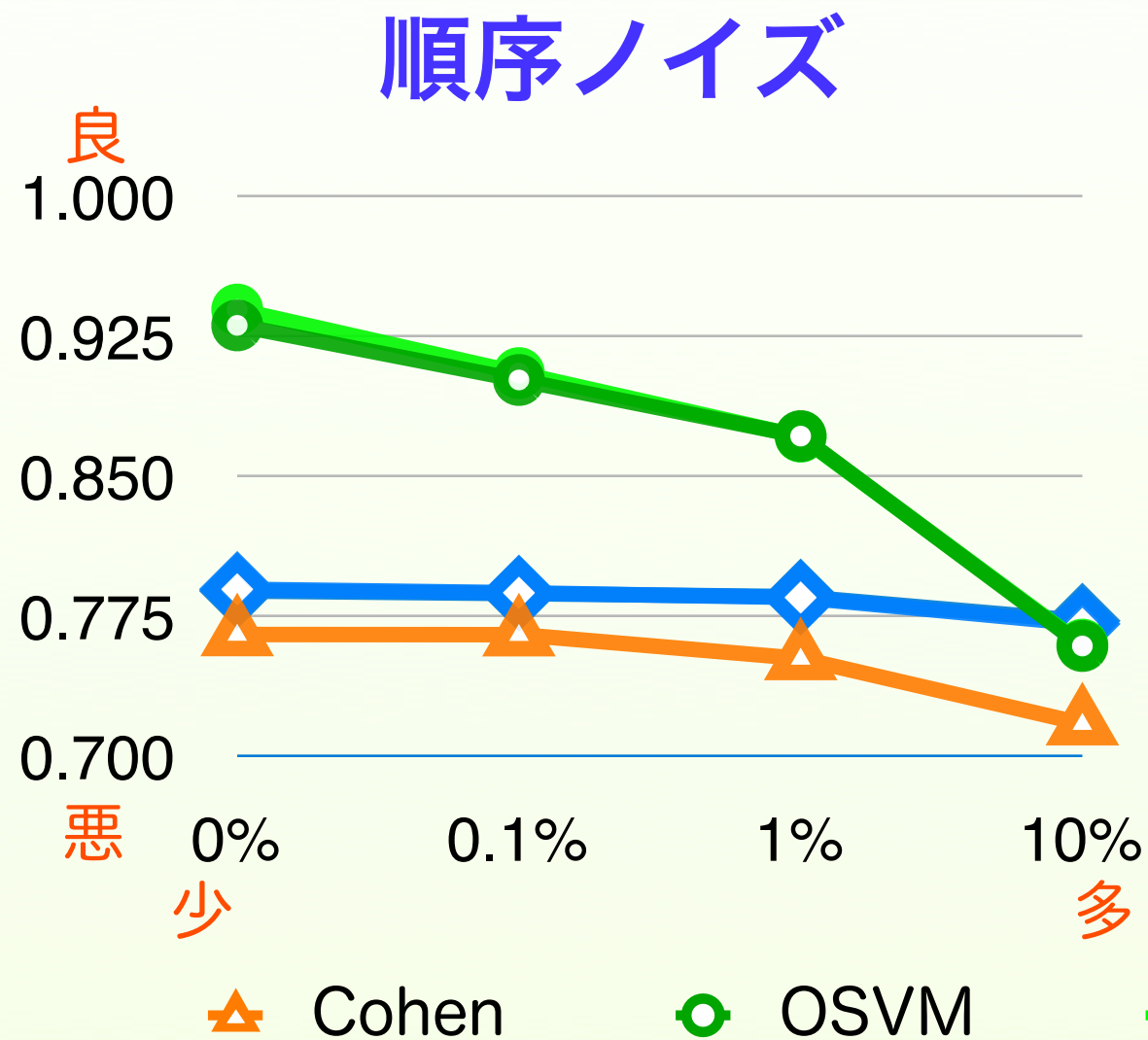
ノイズなし

属性値ベクトルはそのまま $x_i = (x_{i1}, \dots, x_{ik})$

ノイズあり

正規分布ノイズを掛ける $x'_{ij} = x_{ij} \times N(1, \sigma)$

順序ノイズと属性ノイズ (結果)



縦軸：予測精度 横軸：ノイズの量

順序ノイズには非SVM系が頑健

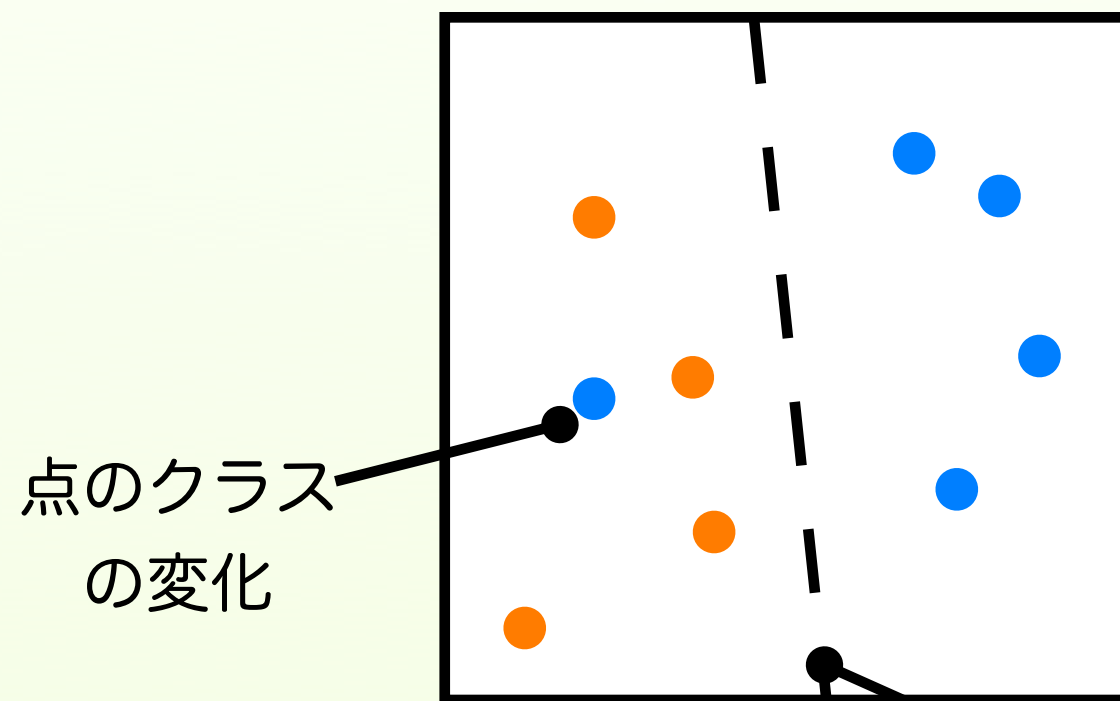
属性ノイズにはSVM系が頑健

SVM系手法とノイズ

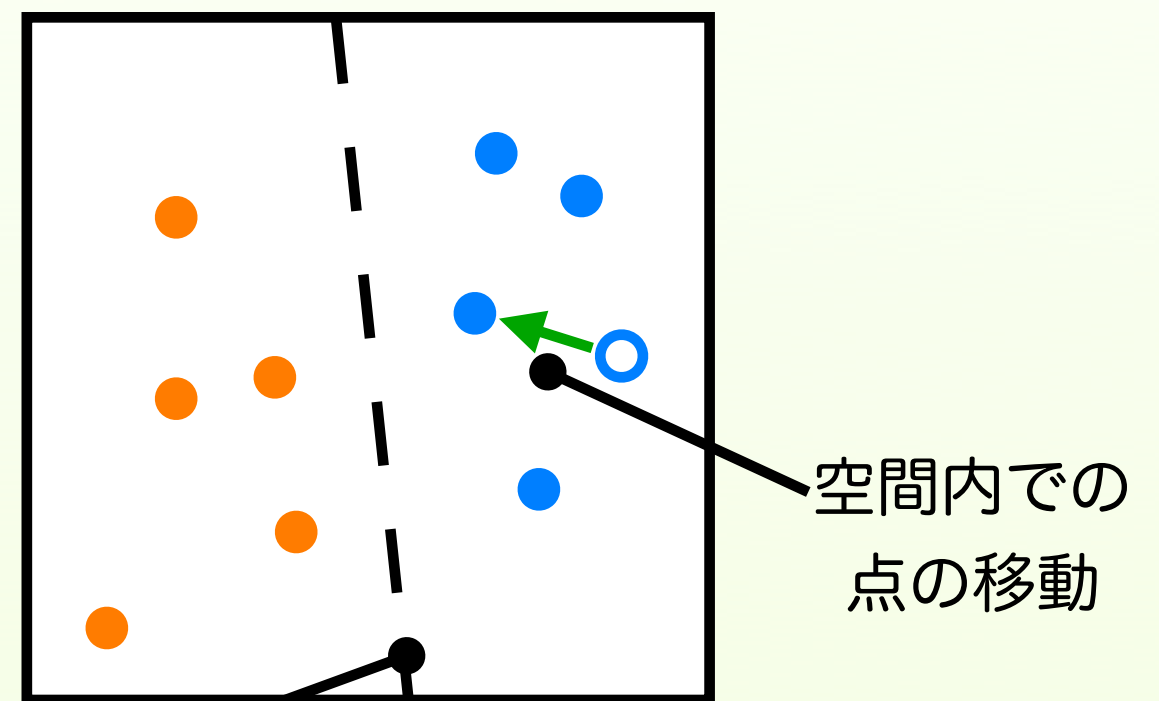
基本的には、判別問題のSVMと同様

OSVM: 第j位より上位になるか？ SVOR: どちらの対象が上位か？

順序ノイズ



属性ノイズ



識別平面

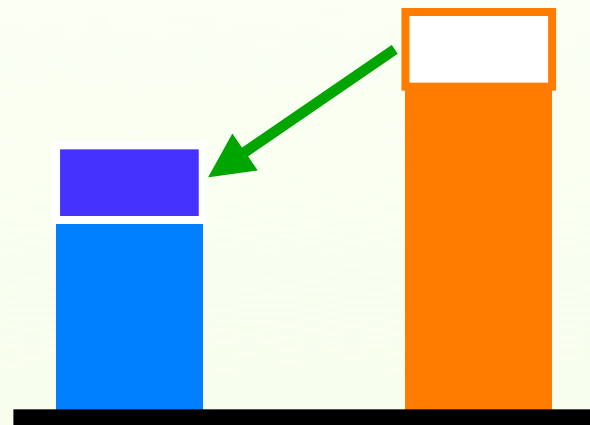
サーポートベクトルになりやすく
ほぼ確実に影響

識別平面を超えなければ影響なし

Thurstone系手法とノイズ

対象対 $A > B$ と $B < A$ 多少が推定精度に影響

順序ノイズ



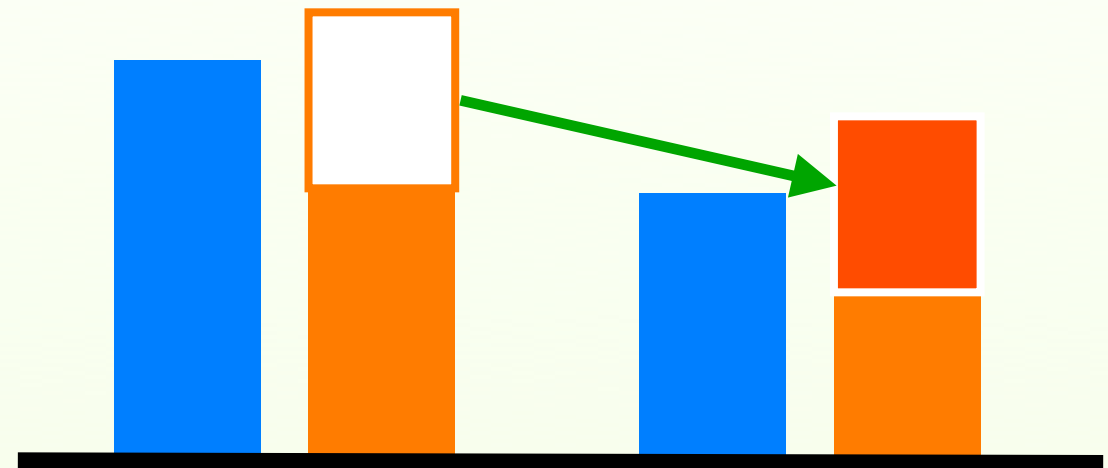
$A \succ B$

$B \succ A$

属性空間中の同じ位置で
 $B > A \rightarrow A < B$ へと事例が移動

この位置で逆転が
なければ結果は不変

属性ノイズ



$A \succ B$ $B \succ A$

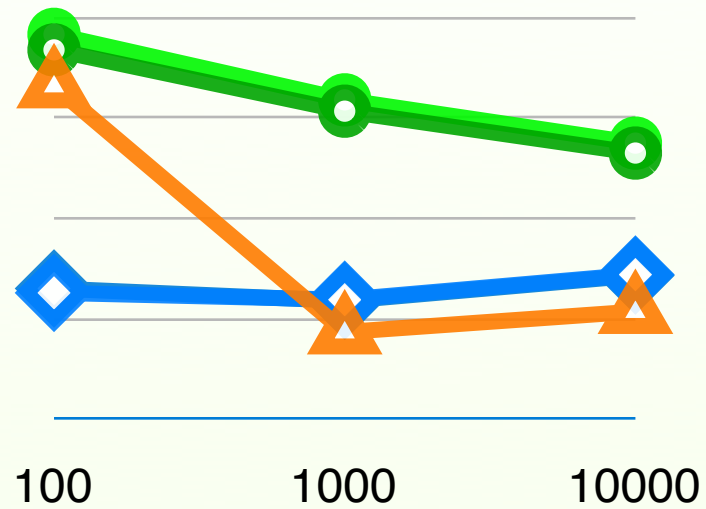
$A \succ B$ $B \succ A$

$B > A$ の事例が属性空間中の
違う位置へと移動

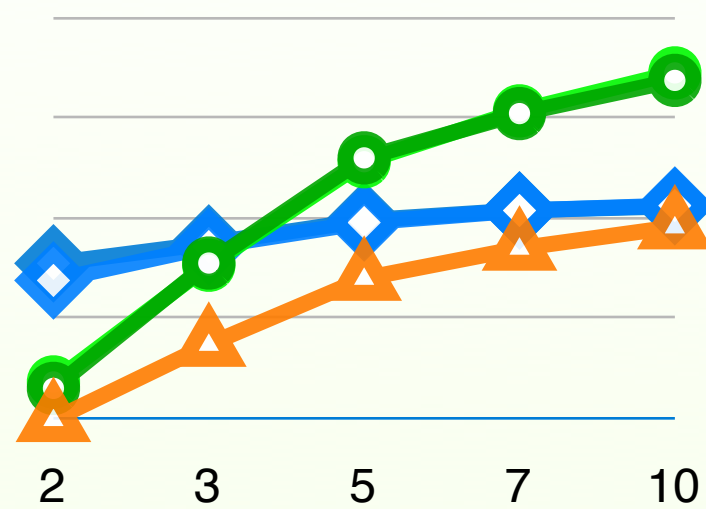
移動元と先の両方で
逆転しないようでない
と結果が変わる

訓練データの量の影響

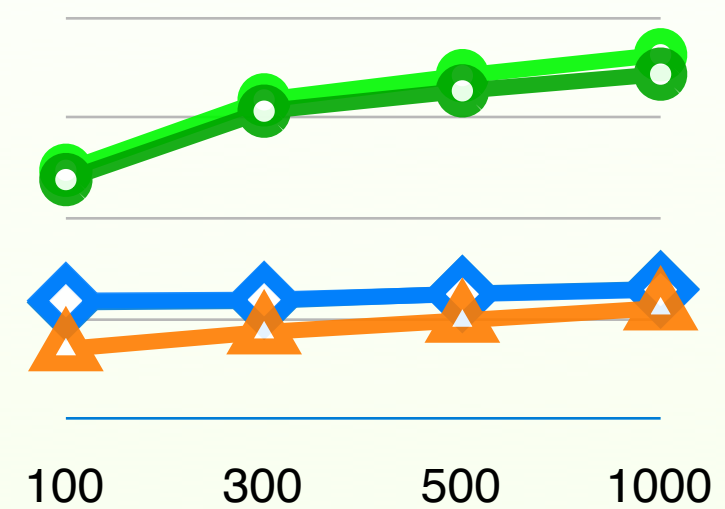
対象の種類数



標本順序の長さ



標本順序の数



△ Cohen

○ OSVM

○ SVOR

◇ TR

◇ ATM

	SVM系 低モデルバイアス	非SVM系 高モデルバイアス
対象の種類数が多い 高い汎化能力が必要	対象数少 過適合	対象数多 適合不足
順序の長さや数が少 情報量が不足	情報量少 推定精度低下	情報量多 適合不足

まとめ (1)

Cohenの方法

- ▶ 予測精度は低いが、最も高速
- ▶ オンライン学習に適す
- ▶ 事後確率関数の学習に低バイアスモデルも利用できるが、計算量は増加する可能性

SVM系 (OSVMとSVOR)

- ▶ 最も低速だが高精度の予測
- ▶ 順序数 $\times$ (順序長) 2 は $10^5 \sim 10^6$ が限界
- ▶ 属性ノイズには頑健だが、順序ノイズには弱い
- ▶ カーネルを変えることで、モデルバイアスを高くもできるが、計算量はそのまま

まとめ (2)

Thurstone系 (TRとATM)

- ▶ 推定精度も学習時間も中程度
- ▶ 標本順序が多くても適用可能だが、
対象の種類数は $10^5 \sim 10^6$ が限界
- ▶ 回帰分析に低バイアスモデルも利用できるが、
計算量は増加する可能性
- ▶ ATMはTRに追加計算が必要だが、精度は変わらないのでTRが
よい

今後の予定

- ▶ 各手法でモデルバイアスを変えて検証
- ▶ 対象の種類数が多い実データでの検証