

飼いならしを用いた 協調タグ付けのタグ予測

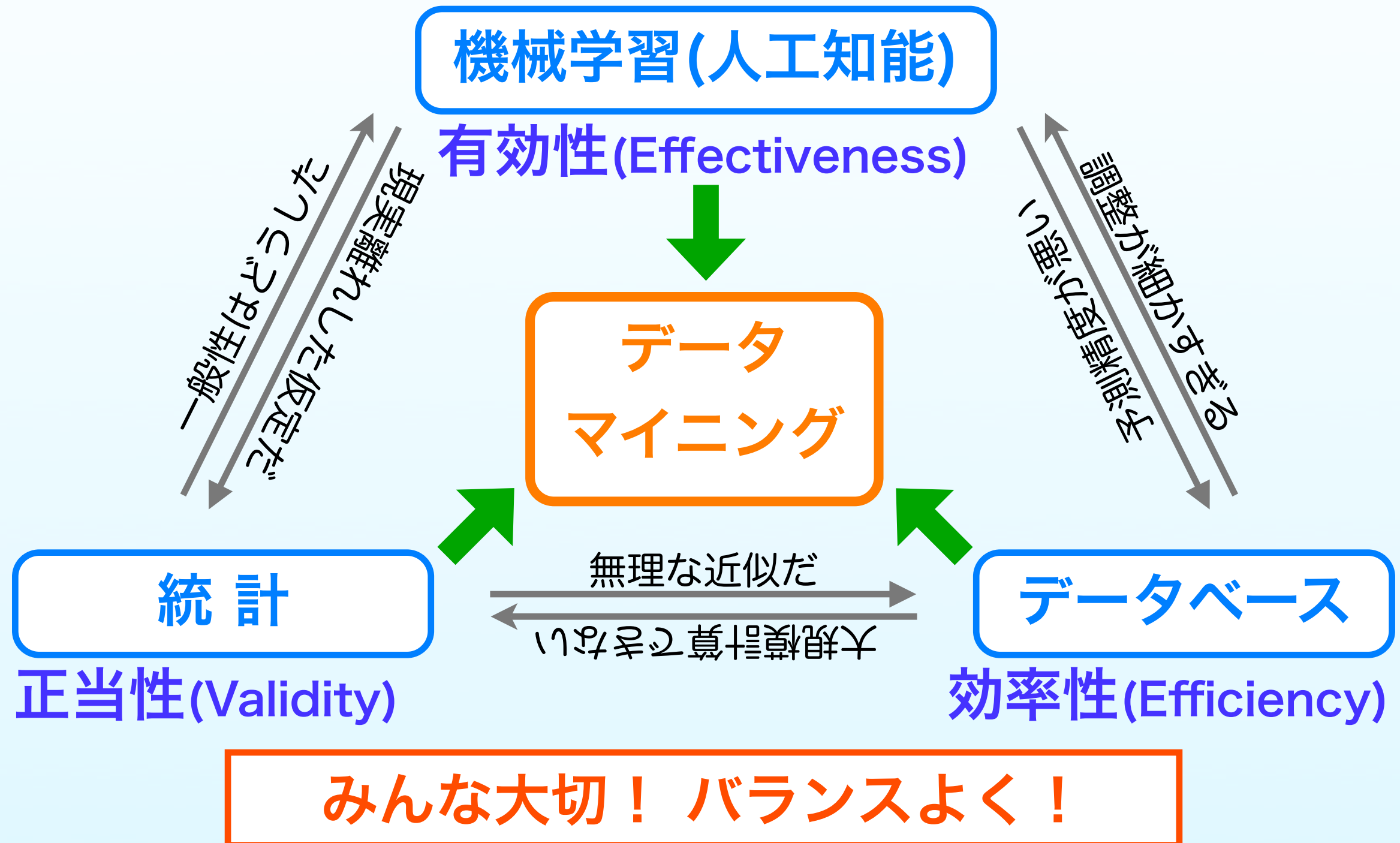
統計関連連合大会 (2008.9.8)

神畠 敏弘, 濱崎 雅弘, 赤穂 昭太郎

産業技術総合研究所

<http://www.kamishima.net/>

データマイニング



Z.-H. Zhou "Book Review: Three Perspectives of Data Mining", Artificial Intelligence, vol.143, (2003)

データマイニング

データ解析技術の探索的な側面(=知識発見)を重視

1. 大規模データ

DBアクセス頻度, メモリ効率, 計算量は線形

2. データ収集が計画的や静的ではない

はずれ値, 欠損値, ストリームデータの処理

3. 新しい種類のデータやパターン

時系列, 木, グラフ, 順序データ, 相関ルール

4. 人間による知識の新規性や有用性の判断

データの可視化, 補助情報や制約の利用

概要

- ◉ 飼いならし学習

信頼できる少数の飼育データと信頼できない多数の野生データから高精度の予測

- ◉ BaggTaming

飼いならし学習問題を扱う手法として, Bagging を拡張したBaggTaming法を開発

- ◉ 協調タグ付け

飼いならし学習の枠組みを利用して, Webページに付加されるタグを高精度で予測する実験

飼いならし

機械学習：一定の基準で与えられた教師情報

教師情報管理の
コストは高い

大量のデータを集
めるのは困難

予測精度
は低い



飼育データ



教師情報は高品質



少量しか利用できない

野生データ



教師情報は低品質



大量に利用できる

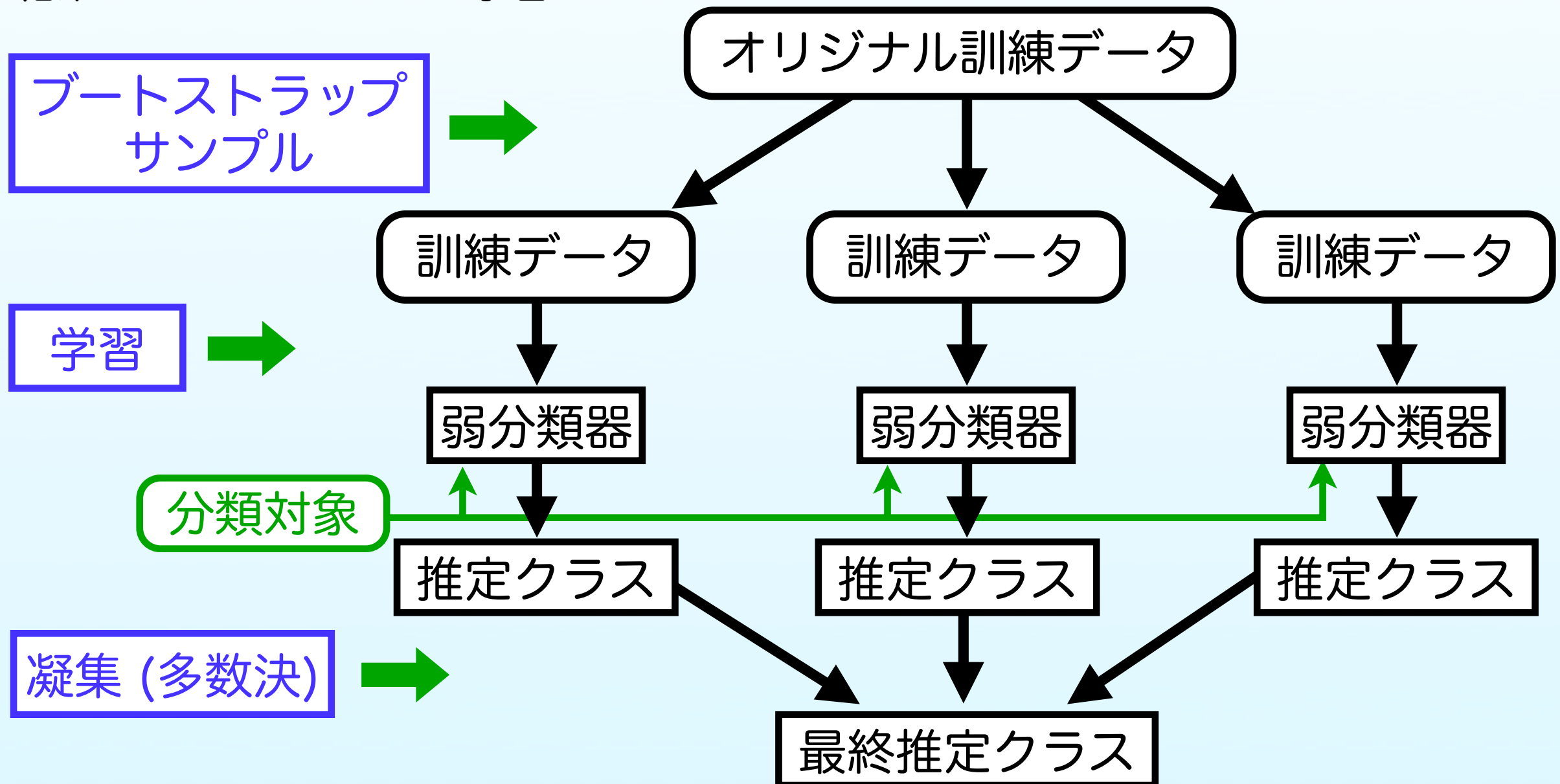
互いの欠点を補うように、両方のデータから学習

飼いならし (Taming)

Bagging

Bagging Bootstrap AGGREGatING

一つの訓練事例集合から複数の分類器を作り，各分類器の分類結果をまとめて最終結果とするアンサンブル学習の一つ



バイアス-バリエーション

バイアス-バリエーション理論

$$\text{汎化誤差} = \text{バイアス} + \text{バリエーション} + \text{ノイズ}$$

- ⌚ バイアス：学習に用いたモデルの表現能力に由来する誤差
- ⌚ バリエーション：学習に用いた訓練集合の内容に由来する誤差
- ⌚ ノイズ：本質的にランダムで解消できない誤差



Baggingでなぜ予測誤差が減少するのか？

- ⌚ バイアス：弱分類器が採用したモデルに依存して決まる
- ⌚ ノイズ：解消できない

弱分類器を多様なサンプルから学習してバリエーションを減らす

BaggTaming

さらにバリエアンスを減らすには？
より多様な訓練事例をサンプリングする



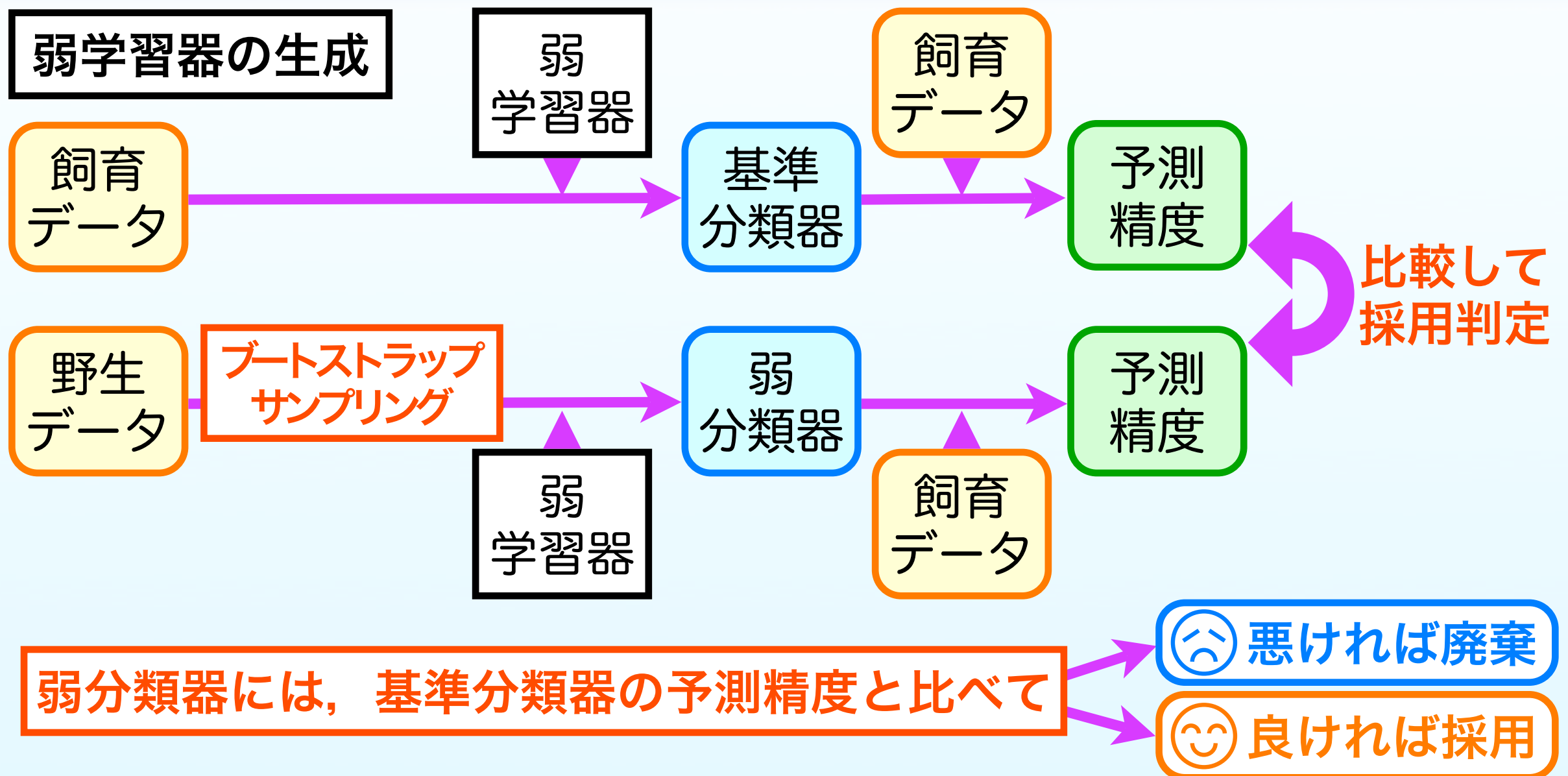
野生データの事例は飼育データよりずっと多様
訓練データを野生データからサンプリング

野生データには不適切な事例もある
訓練データが適切かをチェックし、不適切なら再サンプリング



野生データから学習した学習器が飼育データを
適切に分類できるかどうかでフィルタリング

BaggTamingの弱学習器



- ◉ 予測精度の比較判断にZ検定を採用
- ◉ 採用された弱分類器が一つも無い場合にデフォルト弱分類器を利用

協調タグ付け

協調タグ付け (Collaborative Tagging)

ソーシャルブックマーク

- ◎ 利用者は好きなWebページを登録できる
- ◎ それらのページに関連するキーワードであるタグを付けることが可能
- ◎ Webページとタグを、他の利用者と共有できる

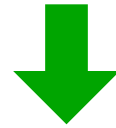


共有したタグは、Webページの検索や整理に利用可能

協調タグ付け

[Golder 06] のタグの不整合の問題

タグの持つ意味やタグの選択基準は利用者によってバラツキがある



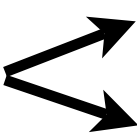
ある利用者のタグは、別の利用者にとっては不適切

抽象性の階層：利用者によって適切な抽象度のレベルは違う

犬のページに「チワワ」「犬」「ペット」タグを付ける人がいる

polysemyな語：利用者は微妙に異なる意味のものを検索してしまう

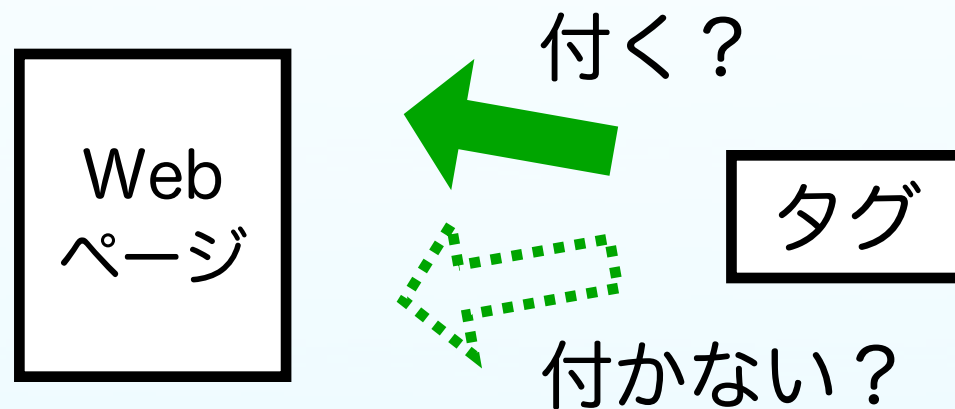
関係のある意味をもつ同形の語

例：window  単なる壁にあいた穴
壁の穴に板ガラスをはめたもの

類義語：同じ対象が別のタグで表される

タグ予測問題

タグ予測問題



任意のWebページに，指定したタグが付加されるべきかどうかを予測

利用法

- ⦿ 新規ページへの予測タグを利用した検索や整理
- ⦿ 利用者がタグ付けするときの候補を示し省力化

クラス分類問題

- ⦿ クラス：指定したタグが付く・付かないの2クラスへの分類
- ⦿ 特徴ベクトル：指定したタグ以外の各タグの数を要素とする

単純ベイズ：クラスが与えられたときの特徴ベクトルの条件付き分布は多項分布である多項モデルを採用

飼いならしによるタグ予測の個人化

特定の個人にとって適切なタグを予測する

従来は対象利用者が過去に付けたタグから学習することで実現

さらに予測精度を向上させるために飼いならしを利用

飼育データ

対象利用者が過去に付けたタグ
タグは個人化された基準によく
従うが、タグ付けされたページ
数は限定されている。



タグ付け基準は明確



少量しかない

野生データ

対象利用者以外が付けたタグ
タグ付けの基準は曖昧だが、多
くの人により多数のタグ付け
データが得られる。



タグ付け基準は曖昧



大量にある

実験設定

- ◉ Webページに付けられた数が**上位20位までのタグを選択**し、これらの各タグが任意のWebページ付加されるかどうかを検証
- ◉ 各タグについて、そのタグを付加した**ページが最も多い利用者のタグ情報を飼育データ**とする
 - ◉ トップ利用者が一人で付けたタグは、その選択基準が安定しているので飼育データとして利用した
- ◉ 各タグについて、タグを付加した**ページ数が上位2～19位の利用者が付けたタグを飼育データ**とする
 - ◉ 19人の野生利用者の間でタグ付け基準の不一致もあるため、飼育データよりタグ付けの基準はあいまい

実験結果

20データに対する実験で、正解率に有意な差があるかで判定した勝敗数

飼育データ数	ALL	1/2	1/4	1/8	1/16
旧手法 (BT/Bagg)	5/2	8/3	8/2	10/2	11/1
新手法 (BT/Bagg)	2/0	6/1	8/1	10/0	9/0

BT=BaggTaming

Bagg = 飼育データのためのBagging

飼育データに対する予測誤差を交差確認で求めた

野生データ数は固定で、飼育データを全部から1/16まで段階的に減らした

- ◉ 飼育データの大きさにかかわらず、BaggTamingと野生データの併用によって予測精度が向上
- ◉ 飼育データが少ない場合に予測精度の向上はより顕著
- ◉ 野生・飼育の両方や野生データのみを用いた学習と比べてもBaggTamingが優れていた
- ◉ 新手法により予測精度が低下する割合が減少
- ◉ 新手法の実行時間は、旧手法の30%ほどになった

帰納転移

帰納転移 (Inductive Transfer)

一つ以上のタスクで学習した知識を，新たなタスクで効果的な仮説を効率的に得るために，維持して適用すること

過去の研究

- ◉ ベイズの枠組みで，複数のタスクに共通する事前分布を獲得
- ◉ 階層型のニューラルネットで複数のタスクに共通する中間層を採用
- ◉ 各タスクに固有のモデルと，全タスクに共通するモデルを考え，これらの混合モデルを考える
- ◉ 補助タスク用の各事例の重み付けと，学習モデルの獲得を交互に繰り返す
- ◉ 補助タスク用データの重みを，主タスク用より小さく設定する
- ◉ 特徴ベクトルの中に主タスク，補助タスク，共通部分を作り込んでおく

帰納転移

補助用の学習データは、他のタスクにとっては整合性のある教師情報がある？

全タスクを同時に解くことに興味があるか、主タスク以外には興味がないか？

	あり	なし
整合	マルチタスク学習	ドメイン適合
不整合		飼いならし

まとめ

まとめ

- ④ 飼いならし学習の提案

信頼できる少数の飼育データと信頼できない多数の野生データから高精度の予測

- ④ BaggTaming法の開発と改良

野生データからサンプリングし，飼育データを使ってその正当性を検証するBagging

- ④ 協調タグ付け問題で有効性を検証

特定の利用者のタグ付けを野生データの併用でより高精度に予測

おまけ

- ④ データマイニングと統計数理研究会：<http://sigdmsm.org/>

- ④ 朱鷺の杜Wiki：<http://ibisforest.org/index.php?FrontPage>

- ④ ホームページ：<http://www.kamishima.net/>