日本語単語分割を題材としたサポートベクタマシンの能動学習の実験的研究

颯々野 学 著

小野寺 喜行

自然言語処理のアプローチ

教師あり学習

- 標準的な手法
- 問題点：大量のタグ付きコーパスが必要

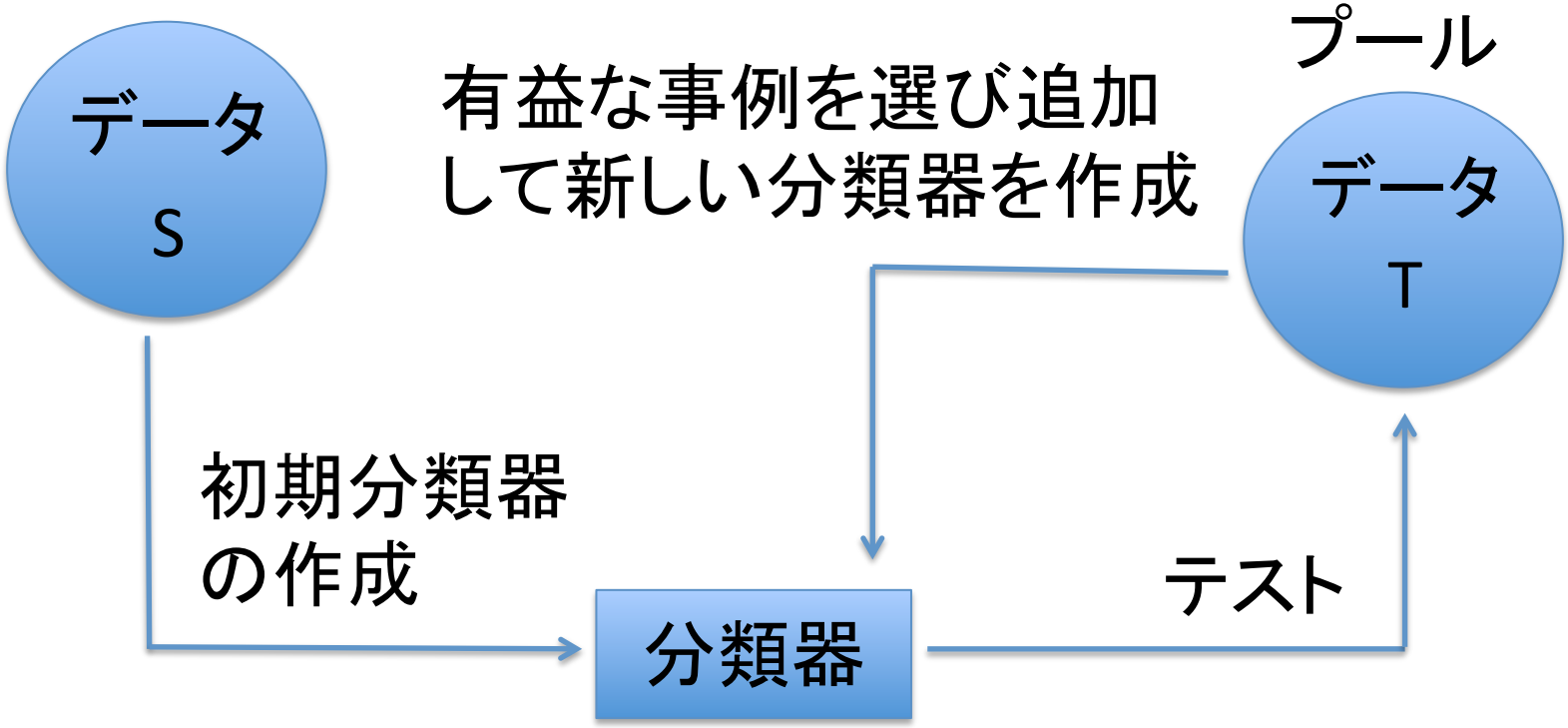
問題点を解決するアプローチとして能動学習を考える

能動学習

分類器がラベル付けすべき事例を選び、人間にラベル付けを要求する

性能を保ったまま、コストを下げることが期待できる

プールに基づく能動学習
（pool-based active learning）
（Lewis and Gale 1994）



事例選択アルゴリズム
(Tong and Koller 2000; Schohn and Cohn 2000)

プール全ての事例x_iについて$f(x_i)$を計算

$$f\left(\boldsymbol{x}\right)=\sum_{i=1}^{l} y_i \alpha_i K\left(\boldsymbol{x}_i, \boldsymbol{x}\right)+b$$

Kはカーネル関数（kernel function）
$b \in \boldsymbol{R}$は閾値　α_iは重み　$y_i \in \{+1,-1\}$

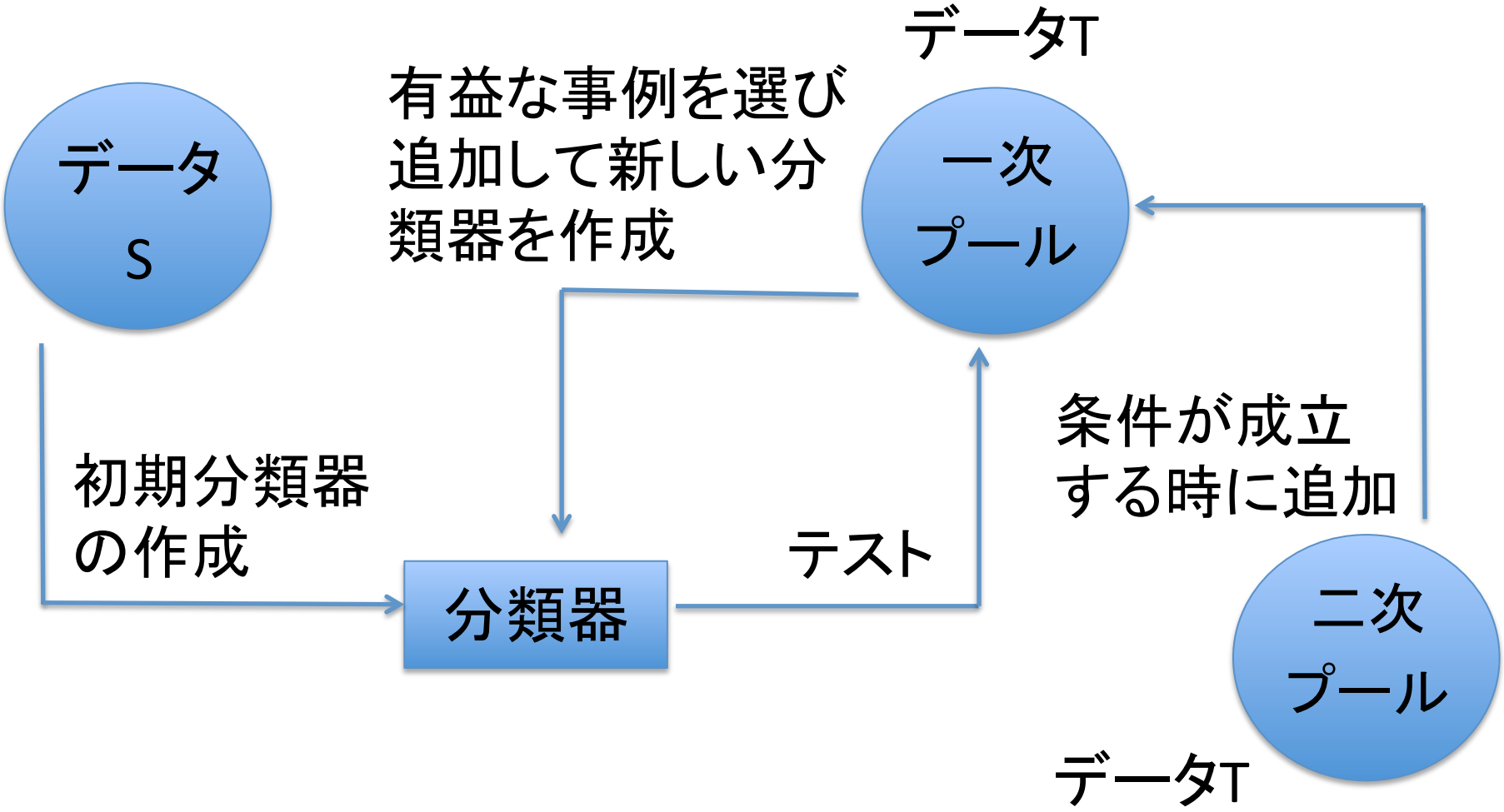
$|f(x_i)|$の小さい順にx_iをソートし、小さい方からm個の事例を選択

プールに基づく能動学習の弱点

能動学習の初期段階で、大きなプールを使った分類器の方が小さな時より多くのラベル付き事例が必要

提案手法：2プールアルゴリズム

2プールアルゴリズム


データT
有益な事例を選び
追加して新しい分
類器を作成
データ
S
一次
プール
条件が成立
する時に追加
初期分類器
の作成
テスト
分類器
二次
プール
データT

一次プールに追加する条件

アルゴリズムA

分類器のサポートベクタ数の増加率が下がったとき

アルゴリズムB

分類器のサポートベクタ数が閾値dを超えたとき

$$d = N\frac{\delta}{100}, 0 < \delta \leq 100$$

δ : 追加するタイミング

N: トレーニングセット中のラベル付き事例数と一次プールのラベルなし事例数の合計

一次プールに追加する事例数

アルゴリズムA、アルゴリズムB共通

一次プールのラベル付き事例数

＋

一次プールのラベルなし事例数

倍になるように追加

実験に使うテストケース

日本語単語分割

日本語の文字の連続: $s = c_1 c_2 \cdots c_m$

単語境界b_i: c_iとc_{i+1}の間に存在

+1（境界である）、-1（境界でない）

日本語単語分割はb_iのクラスを定義する問題

そのクラスの決定にSVMを用いる

文字c_iの属性

属性1 文字種 t_i

以下のいずれかの値

平仮名、片仮名、漢字、数字、英字、漢数字、記号

属性2 文字コード k_i

値の範囲：1～6879

境界b_iのラベル推定

$c_{i-1}, c_i, c_{i+1}, c_{i+2}$の4文字の属性を使う

推定のための素性

$t_{i-1}t_it_{i+1}t_{i+2}$, $t_{i-1}t_it_{i+1}$, $t_{i-1}t_i$, t_{i-1}, $t_it_{i+1}t_{i+2}$, t_it_{i+1}, t_i, $t_{i+1}t_{i+2}$, t_{i+1}, t_{i+2},
$k_{i-1}k_ik_{i+1}k_{i+2}$, $k_{i-1}k_ik_{i+1}$, $k_{i-1}k_i$, k_{i-1}, $k_ik_{i+1}k_{i+2}$, k_ik_{i+1}, k_i, $k_{i+1}k_{i+2}$, k_{i+1}, k_{i+2},

実験

EDR日本語コーパス（EDR 1995）（208000文）から、学習用（20000文）とテスト用（1000文）をランダムに選択

前述の素性を用いて各事例を作成

SVMを用いて実験

実験

実験1 ランダムに選ぶラベル付き事例の数を変化させ、受動学習

結果 事例数が増加 ⇒ 精度が上昇

実験

実験2 プールにある事例の数を変化させ、能動学習（アルゴリズムA）

結果

- 精度96%に達するのに必要なラベル付き事例数が受動学習に比べ70%減
  （2500文のプールの場合）
- 20000文のプールの場合もおおむね同じ割合
- 学習の初期、大きいプールの方が小さいプールよりも精度の上がり方が悪い

実験

実験3 プールにある事例の数を変化させ、能動学習（アルゴリズムB）

結果 精度97%に達するのに必要なラベル付き事例数が受動学習に比べ82.6%減、従来手法に比べ40.7%減

（20000文のプールの場合）