自 然 言 語 処 理　146－13
（２００１．１１．２１）

AdaBoostと能動学習を用いたテキスト分類

塚本浩司 颯々野学

富士通研究所

〒211-8858 神奈川県川崎市中原区上小田中 4-1-1

{tukamoto,sassano}@jp.fujitsu.com

本稿では、テキスト分類のタスクに対し、AdaBoostを用いた能動学習の手法を適用した結果を報告する。対象文書としてReutesr-21578コレクションを用いたところ、能動学習なしのAdaBoostの精度を、約1/10の正解データ量で実現できた。また、文書の表現を語の出現回数を特徴値とするベクトルから、語が出現したかどうかのbinary値を特徴値とするベクトルに変えることにより、能動学習で観察されることのあるサンプル数の増加に伴う精度の劣化について考察した。

Text Categorization using Active Learning with AdaBoost

Koji Tsukamoto and Manabu Sassano

Fujitsu Laboratories, Ltd.,

1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki 211-8588, Japan

{tukamoto,sassano}@jp.fujitsu.com

In this paper, we describe an active learning method with AdaBoost on a text categorization task. This method reduces by as much as one tenth the amount of labeled data needed by AdaBoost without the active learning method for Reuters 21578 newswire articles. We also discuss phenomena that the accuracy of the categorization decreases when the size of training data increases by changing the value of feature vectors from continuous to binary.

1 はじめに

テキスト分類に対するニーズは数多く存在する。Web ページや社内文書が大量に存在する場合、それが予め決められた内容毎に分類されていると、文書を効率的に発見することができる。例えばYahoo!のディレクトリでは、Web 文書を内容ごとに分類し、ユーザビリティの向上をはかっている。また、ヘルプデスクでは、問い合わせメールは一つのアドレスに集中して送信された後に、内容ごとに割り振られることも多い。テキスト分類は、このように文書の内容に基づいて分類することを目的とする。つまり「あるテキストが、予め決められたいくつかのカテゴリのどれに属するかを決める」というタスクである。

しかし、これまで用いられたテキスト分類の技術はかなりの人手コストを必要とするという問題があった。10 年ほど前までは分類ルールや検索式を人手で作成することが主流であった。この方法は、ルール作成者に経験や知識を要求し、カテゴリが変更された場合にもルールを新たに人手で作成する必要がある。そのため、最近は学習による方法にとって変わられつつある。学習に基づく方法は様々提案されている [Sebastiani 1999]。このアプローチの弱点は、分類された正解文書を必要とする点である。人手でルールを作成する場合に比べれば低いとはいえ、予め一定量の正解文書を用意することはコストが高い作業である。機械学習を用いたテキスト分類では、この高コストの作業をできるだけ避けて、より少ない正解文書からどれだけ高い精度が得られるかがポイントとなる。これに対して能動学習 (active learning) と呼ばれる手法が有効であると期待できる。能動学習とは、学習器が学習データを能動的に選択できる手法で、少ない正解事例から高い精度が得られることが知られている。

そこで、我々は能動学習がテキスト分類にどの程度有効かを明らかにすることを目指す。本論文では、Abe が提案した AdaBoost の能動学習手法である Query-By-Boosting [Abe 1998] を新たにテキスト分類に適用し、その振舞いを詳しく調べた。対象としたデータは、Reuters 21578 コーパスである。これは約 10,000 文書の規模を持ち、さまざまなテキスト分類のシステムの評価に用いられているものである。

実験の結果、AdaBoost の能動学習を用いた場合、能動学習を使わないときに比べて、約 1/10 の正解文書の量で同程度の精度が達成できることが分かった。更に、テキストの表現の変化がサンプル数の増加と精度の劣化に与える影響を調べた。テキストの表現を連続値 (語の出現回数) を特徴値とするベクトルの場合と、binary 値 (語が出現したかどうか) を特徴値とするベクトルの場合を比較した。

以下、2 章で AdaBoost と能動学習の枠組を説明し、3 章で他の研究との関連について述べる。4, 5 章で実験の説明および実験結果を示す。最後に 6 章で簡単にまとめる。

2 AdaBoost を用いた能動学習

2.1 AdaBoost

Boosting は PAC 学習モデルの枠組 [Kearns 1994] を応用し、[Schapire 1990] により提案された。これには、予めそこで利用する弱学習器の精度の見積りができなければならないなど、いくつかの実用上の問題が存在したが、[Freund 1997a] により提案された AdaBoost では、この問題は解消され、現実のデータに対してもそれまでの学習器と比べておおむね高い性能が出ることが実験的に示されている [Freund 1996]。

AdaBoost のアルゴリズムを以下に示す。基本的な戦略は、与えられた学習データに対して AdaBoost の用いる弱学習アルゴリズムを繰り返し学習させる際に、間違い易いデータに対してデータの重みを増し、集中的に学習していくことにある。

[Input]
学習データ:$(x_1, y_1), ..., (x_m, y_m)$
$(x_i \in X, y_i \in Y = \{-1, +1\})$
ラウンド数: T

[Learn]
サンプルの重み: $D_1(i) = 1/m$
For $t = 1, ..., T$

1. 弱学習器を分布 D_tに基づいて学習させる。
2. 弱仮説 h_t 及び、そのエラー$\epsilon_t = \Pr_{i \sim D_t}$を求める。
3. h_tの係数 α_tを決める。$\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$
4. 重み D_tを更新する。Z_tは規格化係数。

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases}$$

$$= \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

[Decision]

$$h_f(x) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$$

なお、本稿では弱学習器として、連続値の特徴を最も information gain の高くなる任意の値で分割する決定株 (Decision Stump[Schapire 2000]) を用いた。本稿で用いた決定株の分割ルールは、C4.5[Quinlan 1993] のノードの分割ルールと全く同じである。

2.2 能動学習

能動学習はランダムに収集された正解事例から学習を行なうのではなく、学習器がそれまでに取得した事例に基づいて必要な事例に関して質問することで正解事例を増やす方法である。通常、ランダムに収集された正解事例には冗長性があると考えられることから、この枠組を用いることで学習に必要な正解事例を減らせると考えられる。能動学習を用いる場合、どの学習器を用い、どの事例に対して質問を行なうかが問題となる。質問事例の決め方はいくつか提案されている。複数の学習器の判定結果が割れるサンプルを選ぶ方式 [Seung 1992], [Liere 1997]、単一の学習器の判定結果がもっとも曖昧なサンプルを選ぶ方式 [Lewis 1994], [Schohn 2000]、もっとも予測誤差の小さくなるようなサンプルを選ぶ方法 [Cohn 1996], [Tong 2001], [Roy 2001] 等である。本稿で用いた方法は 2 番目の方法で AdaBoost の出力する確信度がもっとも 0 に近いもの (付与されるラベルがあやふやなもの) からサンプリングを行なう方法である。AdaBoost を用いた能動学習の枠組は [Abe 1998] により提案されている。アルゴリズムを以下に示す。

[Input]

ラベルの不明なサンプルの集合: $S = (x_1, x_2, ..., x_N)$

ラウンド数: T

最初に準備するデータ数:n_0

1 回のサンプリング数: n

[Learn]

ラベルの分かっているサンプルの集合: $S_1 = \emptyset$

For $t = 1, ..., T$

1. S からランダムに n_0 サンプリングし、ラベルを質問して S_1 に加える。
2. S_1から、AdaBoost を用いて学習を行ない、分類ルールを作る。
3. 生成された分類ルールを用いて、S のうち S_1 に入っていない残りのサンプルを判定し、もっとも決定境界 (decision boundary) からの距離の近いものから n 個のサンプルのラベルを質問する。
4. 3 で得られた結果を S_1に加え、2 に戻る。

[Decision]

$t = T$の時の AdaBoost の出力結果を出力とする。

3 関連研究

AdaBoost による能動学習を用いたテキスト分類自体は、現在のところ報告されていない。[Abe 1998] により、Query-by-Boosting が提案されており、そこでは弱学習器として C4.5 および WMP1 が用いられている。本稿では、AdaBoost において非常に単純ではあるが、高い性能の出る決定株を用いた。またこの方法では一様分布を仮定して一定数サンプリングし、そのうちで最も決定境界からの距離が小さくなる点を次の質問を行なう点をとしているが、今回は予め準備されたサンプルすべての中で最も決定境界に近いものを選ぶ様にした。これは、[McCallum 1998], [Lewis 1994], [Tong 2001] らの方法と似ている。近年、AdaBoost と Support Vector Machines (SVM) との関連が指摘されるようになっている。SVM を用いた場合、能動学習は決定境界との距離の近いものを加えていくという heuristics が存在する。線形の SVM による能動学習を用いたテキスト分類の報告 [Schohn 2000], [Tong 2001] では、決定境界が平面になるので、あるラベルのつく領域は 1 つとなる。本稿では文書を特徴語出現の是非 (0/1) とする場合と、出現数 (連続値) として 2 通り用いることにより、最終的に求まる決定境界が特徴空間を二つに分割するかと、3 つ以上の領域に分割するかとを比較した。

4 実験条件

4.1 対象データ

対象としたデータはReuters-21578 コレクションと呼ばれるもので、Reuters 通信社の速報記事を集めた

ものである (http://www.research.att.com/~ lewis/reuters21578.html から取得可能)。このデータのうち、12,902 文書に対して 118 種類の内容に応じたラベルが付与されている。複数のラベルが付与されている文書も存在する。これを ModApte split と呼ばれる分類方法に従い、9603 文書を学習データとし、3299 文書を評価用データとした。ラベルの付与されている文書の数は、ラベルの種類によってかなり差がある。earnings のラベルは 3964 文書に対して付与されているのに対して、crude-oil は 1 文書にしか付与されていない。本稿では、付与されている数の多い 10 のラベル (earn, acq, money-fx, grain crude,trade, interest, ship, wheat, corn) を対象として分類実験を行なった。このセットに対する評価実験は様々なアルゴリズムを用いて行なわれ、詳細な性能比較などもされている [Sebastiani 1999]。特徴は、文書に出現する単語のベクトルを用いた。stemming、不要語処理などは行なっていないが、文字はすべて小文字に統一した。特徴はラベルがつくかどうかという観点で information gain が上位の語について、語の出現回数あるいはその語が出現したかどうかを 1/0 の binary で表現したものとの 2 種類を用いた。

4.2 学習および分類方法

能動学習およびそこで用いる学習方法は、2 章に示した方法を用いた。特徴語数および Boosting 回数は、一律にそれぞれ 100 語、100 回とした。また、複数のラベルが付与される問題を扱う方法はいくつか提案されているが、今回は単純にそのラベルが付与されるかそうでないかの 2 値判定を 10 のラベルそれぞれに対して学習し、分類するという方法をとった。能動学習でのサンプリングはラウンド毎に 10 あるいは 100 サンプルずつ加えている。ラウンド数は 10 サンプルずつ加える場合には 50 回, 100 サンプルずつ加える場合には 20 回で止めている。

5 実験結果

5.1 サンプリング方法の効果

acq および corn について、サンプル数と精度との関係を示す。比較のため、同数をランダムサンプリングしたものを用い、AdaBoost(能動学習なし) および C4.5 でラベルの判定を行なった結果を示す。これらは、それぞれ乱数の初期値を 5 つ用いて正解データ集合として異なったものを使用し (能動学習の時はスタート時のサンプルのみ) 実験を行なっており、ここに示したのは、その平均値である。この 5 回の F-measure のレンジ (最大値 - 最小値) を併せて示す (Figure 1-6)。

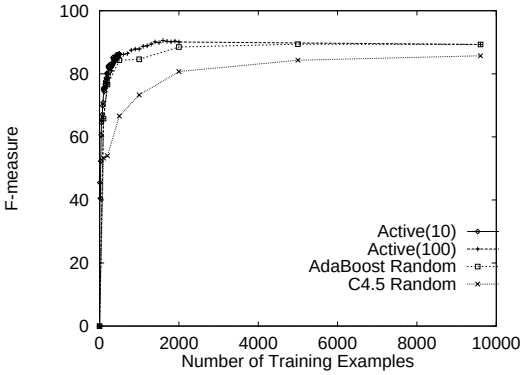

Figure 1: F-measure の文書数依存性 (acq)

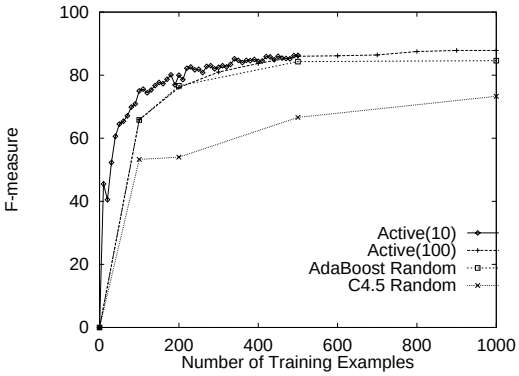

Figure 2: F-measure の文書数依存性 (acq, 拡大図)

acq において、10 ずつサンプルを加えていく場合には、サンプルの数の少ないうちは、ランダムサンプリングにくらべかなり精度が高く、大きな効果のあることが分かる。例えば 100 サンプルから学習した場合の F-measure は、10 ずつサンプルを加えた場合には 75.0 であるのに対し、ランダムサンプリングした場合には 65.8 である。一方、100 ずつサンプルを加えて行く場合には能動学習の効果はそれほど端的にみられない。これは acq のラベルが少数のサンプルからも高い精度を達成し、差が出ないことによると考えられる。実際に、最終的に全サンプルから学習した場合に到達する精度は 89.3 であるのに対

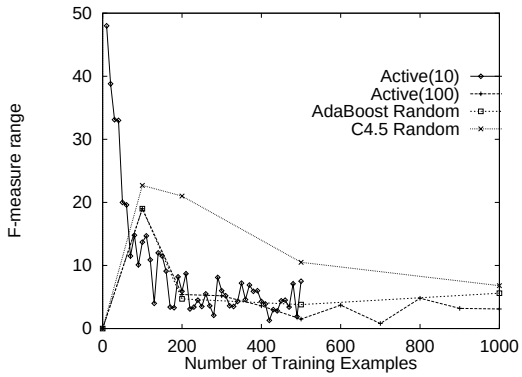

Figure 3: F-measure のレンジ (acq)

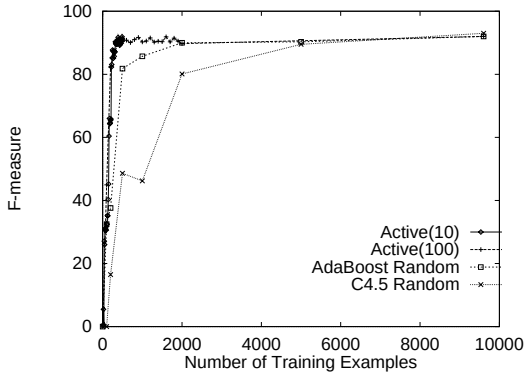

Figure 4: F-measure の文書数依存性 (corn)

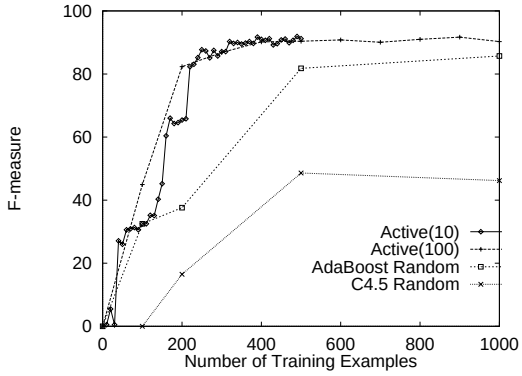

Figure 5: F-measure の文書数依存性 (corn, 拡大図)

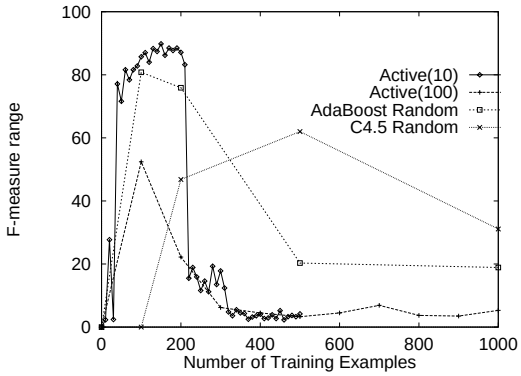

Figure 6: F-measure のレンジ (corn)

し、ランダムサンプリングを行なった場合でも、その 1/20 の 500 サンプルで 84.3 を達成している。ただし、精度のばらつきに関して言えば、100 ずつサンプリングする場合にも、能動学習を用いることで若干小さく押えられることが分かる。

一方、corn においては別の傾向が見られる。この場合、10 ずつサンプルを加えていく場合にも、100 ずつサンプルを加えていく場合にも、初期にはランダムサンプリングと比べ、有意な差がない。サンプル数が 300 くらいから有意な差が生じはじめ、サンプル数が 1000 くらいまでは 10〜20 ポイントのかなり大きな差を保つ。加えていくサンプル数が 10 の場合と 100 の場合では、ほとんど差がない。これは、corn が 10 種類のラベルの中で最も付与率が低く (1.8%)、サンプリングを行なう時点で学習器の作るモデルが、最終的なモデルと大きくずれているので、精度に寄与するサンプルを取れていないことが原因の一つだと考えられる。実際、Figure 6 からも分かる通り、サンプル数の少ない範囲では非常に精度のばらつきが大きく、乱数の初期値によってサンプリングされたものがかなり異なっていることを示している。

精度の信頼度という点では、能動学習はサンプル数が 400 あたりから F-measure のレンジがランダムサンプリングに比べてかなり小さくなっている。これはこの時点で分類に有効なサンプルをほぼとり終っているためだと考えられる。

全体的な傾向を Figure 7, 8 に示す。ラベルのついているサンプルの割合は、1.8〜20% 程度である。この範囲では平均すると、1 ラウンドでのサンプリング数が 100 よりも 10 の方が有効である。精度としては、100 ずつサンプリングした場合には、1000 事

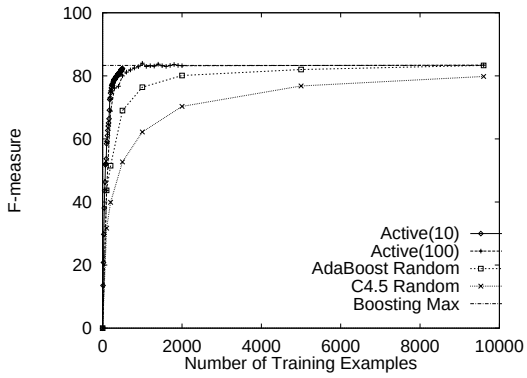

Figure 7: F-measure の文書数依存性 (10 ラベル平均)

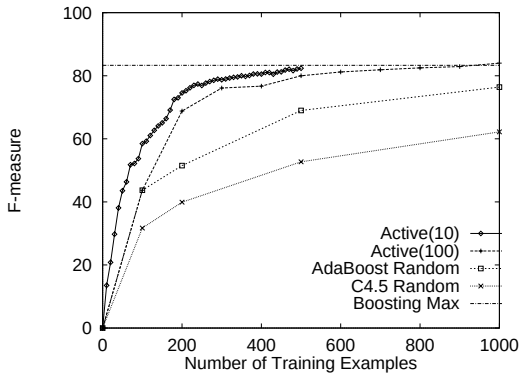

Figure 8: F-measure の文書数依存性 (10 ラベル平均)

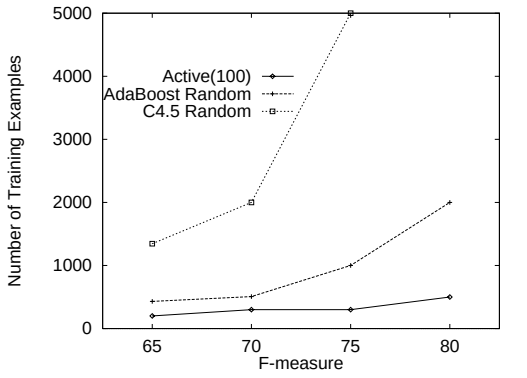

Figure 9: F-measure を達成するのに必要なサンプル数

例のサンプリングをした時点で、9603 文書を用いた場合の精度 (83.4) を初めて上回る。

次に、同じ精度を目指した場合、必要になるサンプル数を Figure 9 に示す。ある精度を達成するために必要なサンプル数はランダムサンプリングを行なう場合には AdaBoost, C4.5 ともに急速に大きくなるが、能動学習を用いた場合、この範囲に限って言えばそれほど急速に大きくなっていない。同等の精度を達成しようとした場合には、能動学習はランダムサンプリングでの AdaBoost の 2 分の 1〜数分の 1 のサンプル数しか必要にならないことが分かる。

5.2 能動学習におけるサンプル数と精度

通常、正解サンプルの数が多い方が学習の精度は良い。しかし能動学習においては、質問を繰り返しサンプル数を増加させて行くと、途中で精度が極大値をとることがある [Lewis 1994], [Schohn 2000]。この節では主に [Schohn 2000] で報告されている現象を参考に検討を行なった。

SVM を用いた [Schohn 2000] の実験では、精度の極大値が特徴的に観察されているのに対し、我々の実験では Large Margin Classifier という点で SVM と類似している AdaBoost を用いたにもかかわらず、極大値は観察できなかった。[Schohn 2000] の実験では SVM の Kernel 関数は線形なので決定境界は特徴空間の中の単純な平面となり、同じラベルだと判定される特徴空間中の領域は 1 つになる (特徴空間は決定境界で 2 つに分割される)。

一方、我々の実験では特徴として単語の出現回数 (連続値) を用いており、AdaBoost で弱学習器として任意の値での分割を許す決定株を用いているので、同じラベルだと判定される領域が 2 つ以上になり得る (例えば、「"oil"という単語が 1 度も出現しないか、3 回以上出現すれば、あるラベルがつく」という様なルールは、特徴空間を 3 つ以上に分割する)。

そこで、AdaBoost で同じラベルだと判断される領域が 1 つになるようにし、[Schohn 2000] で報告された現象を観察できるかを検討した。連続値の特徴を用いたまま AdaBoost が同じラベルのつく領域を 1 つにする適当な方法がなかったので、特徴として単語が出現したかどうかを 0/1 の binary で表現することにした。binary 表現を用いると、AdaBoost の出力結果は特徴空間を平面で 2 つに分割するので

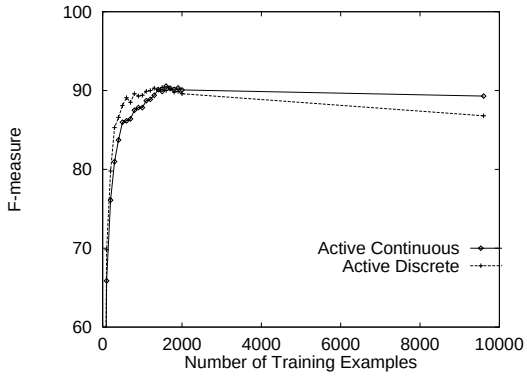

Figure 10: 特徴を変えた場合の F-measure へのサンプル数の影響 (acq)

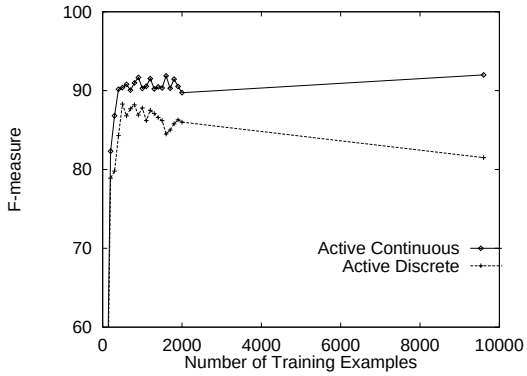

Figure 11: 特徴を変えた場合の F-measure へのサンプル数の影響 (corn)

(Appendix 参照)、同じラベルだと判定される領域は 1 つになる。

Figure 10-12 に実験結果を示す。サンプル数の増加にともなう精度の劣化は、連続値の特徴を用いた場合にはほとんど観察されなかったが、binary の特徴を用いた場合には観察される。これは、複雑なデータを単純なモデルで表現しようとしたために起こったと考えられる。AdaBoost や SVM を用いた能動学習ではサンプリングする点は決定境界の近くである。学習器の構成するモデルが上記の様な平面である場合には、初期にはこのサンプルのラベルによらず、無矛盾なモデルを作れる可能性がある。あるいは矛盾していても、決定境界に近いので影響は少ない。ところが能動学習の後半のフェーズでは、決定境界から遠く離れた、それまでに追加されたサンプルと矛盾するサンプルが追加される可能性が高くなる。このようなサンプルがノイズとなり、学習器の

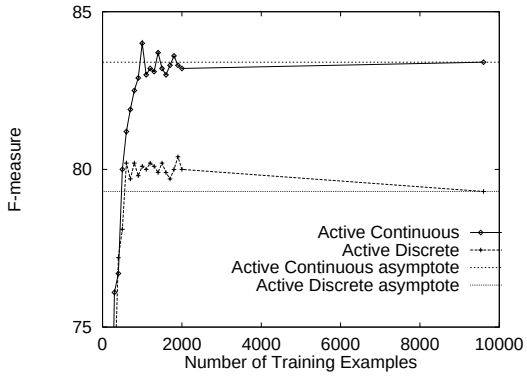

Figure 12: 特徴を変えた場合の F-measure へのサンプル数の影響 (10 カテゴリ平均)

精度を悪化させているものと考えられる。逆に言うと、能動学習で学習できるモデルが実際のモデルに比べて単純な場合には、途中でサンプリングを停止することが有効である可能性がある。

6 おわりに

本論文では、AdaBoost の能動学習を新たにテキスト分類に適用し、能動学習の有効性を詳しく検討した。我々の実験では、Retures 21578 コーパスに対して約 1/10 のデータ量で全ての正解事例を使った場合と同程度の高い精度を得た。更に、特徴の表現が精度の劣化に与える影響を考察した。

Appendix

binary の特徴を用いた AdaBoost の最終出力

$$h_f(x) = \mathrm{sign} \sum_{t=1}^{T} \alpha_t h_t(x)$$

は、以下のようにして特徴空間を平面で分割することが分かる。

$$h_t(x) = \begin{cases} 1 & (x_j = 1) \\ -1 & (x_j = 0) \end{cases}$$

または

$$h_t(x) = \begin{cases} -1 & (x_j = 1) \\ 1 & (x_j = 0) \end{cases}$$

である。新たにβ_tを

$$\beta_t = \begin{cases} \alpha_t & (\text{上側が成立するとき}) \\ -\alpha_t & (\text{下側が成立するとき}) \end{cases}$$

とすることにより、

$$\begin{aligned} h_f(x) &= \sum_{t=1}^{T} \alpha_t h_t(x) \\ &= \sum_{t=1}^{T} \beta_t (2x_{k_t} - 1) \\ &= -\sum_{t=1}^{T} \beta_t + 2\sum_{t=1}^{T} \beta_t x_{k_t} \\ &= -\sum_{t=1}^{T} \beta_t + 2\sum_{j=1}^{J} \{x_j \sum_{t=1,k_t=j}^{T} \beta_t\} \end{aligned}$$

(ただし、x_jは特徴ベクトルのj番目の特徴値、k_tは Boosting の t 回目のラウンドで使われた特徴の ID で、1〜Jのうちの一つ)
決定境界が一枚の平面となり、特徴空間は 2 つに分割される。

Reference

[Abe 1998] Abe, N. & Mamitsuka,H. (1998) *Query Learning Strategies using Boosting and Bagging.* In Machine Learning: Proc. of the 15th Int. Conf.

[Cohn 1996] Cohn, D.A, Ghahramani, Z. & Jordan, M.I. (1996) *Active Learning with Statistical Models.* Journal of Artificial Intelligence Research, 4.

[Freund 1996] Freund, Y. & Schapire, R.E. (1996) *Experiments with a new boosting algorithm.* In Machine Learning: Proc. of the 13th Int. Conf.

[Freund 1997a] Freund, Y. & Schapire, R.E. (1997) *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of Computer and System Sciences, 55.

[Freund 1997b] Freund, Y., Seung, H., Shamir, E. & Tishby, N. (1997). *Selective Sampling usig the Query By Committee algorithm.* Machine Learning, 28.

[Kearns 1994] Kearns, M.J. & Vazirani, U.V. (1994) *An Introduction to Computational Learning Theory.* MIT Press.

[Lewis 1994] Lewis, D.D. & Gale, W.A. (1994) *A Sequential Algorithm for Training Text Classifiers.* In Proc. of 17th Ann. Int. ACM-SIGIR Conf. on Research and Development in Information Retrieval.

[Liere 1997] Liere, R. & Taepalli, P. (1997) *Active Learning with Committees for Text Categorization.* Proc. of the Natioral Conference in Artificial Intelligence.

[McCallum 1998] MaCallum, A.K. & Nigam, K. (1998) *Employing EM and Pool-Based Active Learning for Text Classification.* In Machine Learning: Proc. of the 15th Int. Conf.

[Quinlan 1993] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

[Roy 2001] Roy, N., & McCallum, A. (2001) *Toward Optimal Active Learning through Sampling Estimation of Error Reduction.* In Proc. of the 18th Int. Conf. on Machine Learning.

[Schapire 1990] Schapire, R.E. (1990) *The Strength Of Weak Learnability.* Machine Learning, 5.

[Schapire 2000] Schapire, R.E., & Singer, Y. (2000) *Boostexter: A Boosting-based System for Text Categorization.* Machine Learning, 39.

[Sebastiani 1999] Sebastiani, F. (1999) *Machine Learning in Automated Text Categorization.* Tech.Rep. IEI-B4-31-1999, Consiglio Nazionale delle Ricerche, Pisa, Italy.

[Seung 1992] Seung, H.S., Opper, M. & Somopolinsky, H. (1992). *Query by Committee.* In Proc. of 5th Ann. Workshop on Computational Learning Theory.

[Schohn 2000] Schohn, G. & Cohn, D. (2000) *Less is More: Active Learning with Support Vector Machines.* In Proc. of the 17th Int. Conf. on Machine Learning.

[Tong 2001] Tong, S. & Koller, D. (2001) *Support Vector Machine Active Learning with Applications to Text Classification.* Machine Learning to appear.