決定リストと期待損失を用いた同音異義語識別規則の能動学習

紺野憲一
茨城大学大学院
理工学研究科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

本論文では Roy らが提案した期待損失を利用した能動学習の手法 [1] を同音異義語識別規則の能動学習に応用する。パフォーマンスがあまり高くない機械学習手法を用いても期待損失を利用した能動学習の手法が通常の能動学習の手法 (Query By Committee[2]、以下 QBC と略す) よりも良い結果を出せるのかどうかを、確認することを目的とする。

能動学習は訓練データ構築のコストを低減できることから有用である。能動学習の手法としては QBC が一般的だが、Roys らは期待損失を利用した手法を試み、大きな成果をあげている。ただし、Roys らの能動学習では問題の対象が文書分類であり、ベースとなる機械学習手法は Naive Bayes である。しかも Naive Bayes に bagging の手法も組み入れている。文書分類問題に対しては Naive Bayes がよい結果を出すことが知られており、彼らの手法がうまくいったのは、ベースとなる機械学習手法のパフォーマンスが高かったためとも考えられる。

しかし現実の問題ではどのような機械学習手法を利用してもパフォーマンスの低い規則しか学習できないケースもある。このためパフォーマンスのあまり高くない学習手法を用いる場合でも、期待損失を利用した能動学習の手法が QBC よりもよい結果が得られるのかどうかは調べておく必要がある。

ここでは問題として同音異義語の識別問題を扱う。またベースとなる機械学習手法として決定リストを用いる。期待損失を利用した能動学習に決定リストを利用する場合には、いくつかの実装上の工夫が必要である。1 つは逐次的な学習を実現することである。期待損失を利用した能動学習の手法は、単純に実現すると計算コストが高いため、逐次的な学習が可能であるような学習手法を利用する必要がある。2 つ目は決定リストを作成する際に低頻度の証拠に対する識別クラスの確率を付与しなくてはならないことである。ここではスムージングの手法を採り入れることで対処する。

2 能動学習

能動学習とは学習者が能動的に学習データを選択できる機械学習法である。能動学習はその形態により、データ効率性・計算効率性が高いという利点がある。

膨大なデータがあり、どのデータを学習データにするかを能動学習に選択・学習させることで、少量のデータから十分な精度での効率のよい学習が可能となる。その時どのデータを学習に用いるトレーニングデータとして選ぶかを学習させるのが能動学習である。

2.1 QBC

代表的な能動学習手法として QBC がある [2]。

QBC は Gibbs アルゴリズムを下位学習アルゴリズムにすることを前提とする。Gibbs アルゴリズムの多数のコピーに同一のトレーニングデータを与えて、識別規則を生成する。テストデータをそれら識別規則に与えて、識別結果がもっとも分かれるテスト事例を次にトレーニングデータに追加すべき事例とする手法である。

ただし現実には多くの下位学習アルゴリズムは決定的であり、Gibbs アルゴリズムではない。その点を改良したアルゴリズムとして Query By Bagging という手法がある [3]。これはトレーニングデータからランダムサンプリングを複数回行なう。それぞれのトレーニングデータに対して通常の学習アルゴリズムから識別規則を生成する。テストデータをそれら識別規則に与えて、識別結果がもっとも分かれるテスト事例を次

にトレーニングデータに追加すべき事例とする手法である。

ここでは Query By Bagging も QBC の 1 種として位置付けて、QBC と呼ぶ。

2.2 期待損失の利用

期待損失とは学習によって得られた識別規則を用いて識別を行なった時にどの程度間違えるかを予測した値である。確率分布で表現され識別規則 $\hat{P}_D$ の期待損失 $E_{\hat{P}_D}$ は以下の式で表される。

$$E_{\hat{P}_D} = \frac{1}{|\mathsf{P}|}\sum_{x\in P}\left(1 - max_{y\in \mathsf{y}}\hat{P}_D(y|x)\right)$$

P はテストデータ全体を表し、$\hat{P}_D(y|x)$ は x が y と識別される確率である。

サンプルデータの選択にこの期待損失を用いる。

実際には膨大なデータを大量のテストデータと少量のトレーニングデータに分ける。トレーニングデータから導き出された識別規則を用いてテストデータの判定を行う。次にテストデータから 1 事例を取り出し、先の識別規則で判定されたラベルをつけてそのテストデータをトレーニングデータに追加する。この 1 事例追加されたトレーニングデータに対して識別規則を学習し、その期待損失を求める。これをテストデータの全事例に対して行ない、最も期待損失が小さくなるテスト事例を求める。これが期待損失を利用した能動学習における追加事例の選択方法である。追加する事例が決まった後は、実際にその事例に真のラベルを手作業つけてトレーニングデータに追加する。

3 決定リスト

3.1 決定リストによる実装

決定リストとはある事例の判定を行わせる際、ある判定結果と同時に起こる（共起する）ものを証拠とし、その証拠があらわれた時にその判定結果となる可能性 (予測力) の高い順に並べたリストである。

決定リストの作成は以下の手順で行われる。

ある要素 A がクラス $c_1, \cdots, c_n$ のどのクラスに含まれるかを判定する事とする。

証拠	予測力	クラス
M	2.54234	a
N	2.34272	b
S	2.17457	a
T	1.63453	c
R	1.63387	b
•••	•••	•••
••	••	••
•	•	•
Default	0.2331	a

表 1: 決定リスト

step 1

証拠 evd を設定する。

step 2

クラス c_i と証拠 evd_j とが共起する頻度 $frq(c_i, evd_j)$ をトレーニングデータから得る。

step 3

証拠 evd_j が生じている場合要素 A がクラス c_i である予測力 $est(c_i, evd_j)$ は通常対数尤度比を利用するが、ここでは期待損失の計算に確率を用いる為、予測力は以下で定義する。

$$est(c_i, evd_i) = \frac{frq(c_i, evd_j)}{sum}$$

$$※\ sum = \sum_i frq(c_i, evd_j)$$

また、Default と言う特別な証拠を設定し、リスト内の全ての証拠が見つからなかったときや予測力が Default で設定した値よりも低い場合にはこの Default で指定したクラスとして判定する。

Default に対する予測力は以下のように計算する。

$$Default = \frac{max(|c_i|)}{|D|}$$

D はトレーニングデータ全体である。

step 4

これらを予測率の順に並べるたものが決定リストである。

実際の識別では、このリストの上から見ていき、その証拠が識別する対象事例に含まれるかを調べる。対象事例に証拠が含まれていれば対象事例がその証拠の

指し示すクラスであると考える。ここで決定リストの最後まで証拠が見つからなかった場合は Default クラスを採用する。

3.2 逐次的な学習

期待損失を利用した能動学習を単純に実装すると、計算コストが膨大である。そのため逐次的な処理を行い計算コスト低下を図る。

ここでの能動学習では次のサンプルデータの選択の為にテストデータから 1 事例取り出し、それをトレーニングデータに加え、決定リストを作成する。ここで 1 事例を加える前のトレーニングデータから作成される決定リストと 1 事例と加えた後のトレーニングデータから作成される決定リストはほとんど同じである。

そこで事例を加える前の決定リストを利用して、事例と加えた後の決定リストを作成する。単純に決定リストを作成するもとになる頻度表とそのソート結果を保持しておき、追加された証拠の部分に対してだけ、頻度を更新させ、その部分だけソートし直せば良い。この工夫によりこの追加した部分のみを再計算する方法を用いて大幅な計算量の低下が行える。

3.3 低頻度証拠に対する確率付与

決定リストの場合間引き処理が行われるため低頻度の証拠は決定リストに反映されない。しかし、ここでの能動学習では 1 事例追加した後に低頻度の証拠を省くと、事例を追加する前の決定リストと変化がない。そこで低頻度の証拠も決定リストに反映させるために、スムージングの手法を適用する。具体的には以下の式で $\alpha = 1$ とおく。

$$est(c_i, evd_i) = \frac{frq(c_i, evd_j)}{sum + \alpha}$$

$$Default = \frac{max(|c_i|)}{|D| + \alpha}$$

4 実験

ここでは同音異義語に対して、能動学習の手法を適用してみる。同音異義語とは同じ発音ではあるが違う意味を持った単語の事である。

繰り返し回数	QBC	期待損失
0	0.8409	0.8409
1	0.8268	0.8513
2	0.8452	0.8449
3	0.8563	0.8510
4	0.8643	0.8496
5	0.8578	0.8540
6	0.8667	0.8536
7	0.8692	0.8529
8	0.8711	0.8604
9	0.8740	0.8597
10	0.8755	0.8586
11	0.8758	0.8604
12	0.8829	0.8598
13	0.8905	0.8612
14	0.8919	0.8621
15	0.8923	0.8612
16	0.8935	0.8632
17	0.8954	0.8664
18	0.8969	0.8663
19	0.9020	0.8677

表 2: 決定リスト

ほとんどの場合この同音異義後の識別は容易であるが、人間には識別容易でも機械では識別が難しい単語がいくつか発見されている。「開放」と「解放」そのうちの一つとされている。この「開放」と「解放」に対して判別問題に対し決定リストと QBC を用いた能動学習を行なった。

トレーニングデータとして用いたのは正解が「解放」である５０事例、正解が「開放」である５０事例のあわせて１００事例である。テストデータは６４８９事例である。それぞれ学習によって得た結果から期待損失が低いテスト事例上位１０事例を取り出し、それらをトレーニングデータに加えて新しいトレーニングデータを作成する。

新しいトレーニングデータを用いてまた能動学習を行なうという手順を繰り返す。学習後のトレーニングデータから作られた決定リストでテストデータに対する正解率がどの程度の変化するのかを調べた。以下はその結果である。

5　考察

以下の図は学習による正解率の変化で、図中の折れ線は上から QBC, 期待損失、そして能動学習を行なわずにランダムに事例を加えていったものを表す。

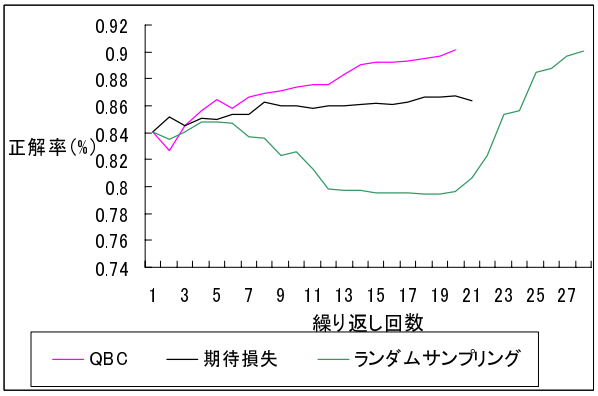

図１：学習による正解率の変化

ランダムに事例を加えていった結果を見ると正解率の低下が見られた。これは追加する事例が学習に対してプラスになるかマイナスになるかわからない事を示唆している。今回のように大きくマイナスに働く事もあるので、ランダムな事例の抽出は行なわない方が良い。

QBC と期待損失とを比べると QBC の方がよい結果となった。また、今回の実験では期待損失による学習の精度向上があまり見られなかった。

期待損失による精度向上があまり見られなかった理由としては、トレーニングデータの事例数が少ない場合に期待損失の値がうまく働いていないことが原因であると思われる。

事例数が少ない場合、低頻度の証拠や証拠の偏りが大きいので期待損失の値をうまくとれない為、最初からある程度の精度の保証がないと期待損失による学習の効果は現れないのではないかと思われる。

最初に精度を保証することは難しいので、事例数が少ない場合にも有効な方法を考える必要がある。

また、決定リストと期待損失の相性が悪い可能性を考慮して、決定リスト以外の判別規則を用いて期待損失を計り、その有効性を検証することも課題となる。

6　終わりに

本論文では期待損失を用いた能動学習を同音異義語識別規則の能動学習に適応させた。

通常用いられる手法に QBC があるが、パフォーマンスがあまり高くない手法でもそれよりよい結果が出せるのかを確認することが目的である。

今回は同音異義語の識別規則の作成に決定リストを用いた。

それを QBC の結果と比較してみた結果, あまりいい結果はでなかった。

原因は小量の事例から求められる期待損失の値があまり適切に働いていないことである。

期待損失の値が有効に働くような判別規則の作成が今後の検討課題である。

参考文献

[1] Nicholas Roy and Andrew McCallum: “Toward Optimal Active Learning through Sampling Estimation of Error Reduction”, Proc. 18th International Conf. on Machine Learning, pp.441–448 (2001).

[2] H. S. Seung and M. Opper and H. Sompolinsky: “Query by committee”, 5th annual workshop on Computational Learning Theory (COLT-92), pp.287–294 (1992).

[3] Naoki Abe and Hiroshi Mamitsuka: “Query learning strategies using boosting and bagging”, International Conference on Machine Learning (ICML-98), (1998).