自 然 言 語 処 理 152－9
（2002．11．12）

情報検索手法を利用した語義判別問題の高速解法

新納浩幸　　佐々木稔

茨城大学工学部システム工学科　　茨城大学工学部情報工学科
shinnou@dse.ibaraki.ac.jp　　sasaki@cis.ibaraki.ac.jp

自然言語処理の個々の問題を分類問題として定式化し，帰納学習の手法を用いて解決するというアプローチは，大きな成功をおさめている．問題に対する精度は，学習により得られる分類規則の精度に依存する．そして分類規則の精度を高めるために，近年，アンサンブル学習が研究されている．アンサンブル学習とは，複数の学習手法を組み合わせ，それらの判別結果を総合して最終的な判別を行う手法であり，単独の学習手法よりも高い精度が得られる．しかしアンサンブル学習には，学習時間と判別時間の点で問題があり，ある程度の精度を出し，しかも高速な学習と判別が可能な帰納学習手法が望まれる．そこで本論文では情報検索手法を利用した分類手法を提案する．k-最近傍法と類似の手法であり，学習時間をほとんど要しない．SENSEVAL2 の辞書タスクによる実験では，本手法は決定リストや Naive Bayes と同程度の精度を出し，判別時間も同程度であった．一方，学習時間に関しては 5 ～ 10 倍高速であった．また決定リスト，Naive Bayes 及び本手法によるアンサンブル学習により，個々の学習手法の判別精度を上回る精度が得られた．

Fast method of word sense disambiguation using information retrieval technique

Hiroyuki Shinnou
Department of Systems Engineering,
Ibaraki University
shinnou@dse.ibaraki.ac.jp

Minoru Sasaki
Department of Computer and
Information Sciences, Ibaraki University
sasaki@cis.ibaraki.ac.jp

Many problems in natural language processing can be converted into classification problems, and be solved by an inductive learning method. This strategy has been very successful. Accuracy for the target problem depends on the learned classifier, so an effective learning method is desired. Against this background, ensemble learning is actively researched. Ensemble learning combines some learning methods, and outperforms each learning method. However, ensemble learning has the problem that it needs much learning time and much classifying time. Therefore, a fast method in learning and classifying is desired. In this paper, we propose the fast learning method which uses information retrieval method. The proposed method is similar to k Nearest Neighbor, and so need little learning time. In experiment, we applied our method to Japanese dictionary task of SENSEVAL2. Precision and classifying time of our method was as the same level as the decision list and Naive Bayes. On the other hand, leaning time was 5 - 10 times faster than them. Moreover, ensemble learning composed of the decision list, Naive Bayes and our method outperformed each method.

1 はじめに

本論文では情報検索の手法を利用した分類手法を提案する．本手法は帰納学習の手法の1つともとらえることができる．その場合，高速な学習が可能という特徴を有し，アンサンブル学習に適している．

自然言語処理の個々の問題を分類問題として定式化し，帰納学習の手法を用いて解決するというアプローチは，大きな成功をおさめている．このアプローチでは，より精度の高い規則を学習できる手法を用いることで精度向上が行える．近年では Maximum Entropy 法や Support Vector Machine を用いた研究が良い結果を出している．これは単一の学習手法を改善してゆく方向とみなせる．一方，この方向とは別に，複数の既存の学習手法を組み合わせることで，分類器の精度を向上させようとする試みも活発に研究されている [1]．複数の学習手法を組み合わせるアプローチには，様々な呼び名があるが，ここではアンサンブル学習と呼ぶことにする．アンサンブル学習では複数の学習手法を準備し，対象の分類問題に対してそれぞれの学習手法を用いて分類器を学習する．テストデータに対しては，作られた複数の分類器の出した判別結果を総合して最終的な判別を行う．これは，個々の学習手法の弱い部分を補う形になるために，得られる判別精度は個々の学習手法単独の判別精度よりも高くなる [6]．

しかしアンサンブル学習では学習に多大な時間がかかるという問題がある．これは複数の学習手法それぞれを用いて学習を行うために避けられない問題である．またテストデータに対する判別でも，複数の分類器で判別を行わなければならないために，通常よりも処理時間を要する．この問題の単純な対応策として，学習や判別を高速に行える学習手法を用いることがあげられる．本論文では，そのような手法の1つとして，情報検索手法を利用した分類手法を提案する．

提案する手法では，訓練データから作られる素性の集合をクラスごとに集める．各素性を単語と見なし，集められた素性の集合を文書と見なす．これにより分類先のクラスの数だけ文書が作られる．テストデータも素性の集合であり，同様に1つの文書と見なせる．ここではテストデータをクエリと見なす．以上より，分類問題はクエリに適合する文書を発見する文書検索の問題と見なせる．このため文書検索の手法を利用して分類問題の解決が可能となる．本手法では，学習に相当する処理が素性ベクトルをクラスごとにまとめるだけであり，学習時間をほとんど要しない．判別に要する時間はクラス数に比例するために，他の学習手法よりも遅い危険性がある．しかし多くの分類問題では分類先のクラスの数がそれほど多くはなく，実際には問題になることは少ない．本論文では語義判別問題を扱ったが，そこでのクラスの数はせいぜい 10 程度であり，判別時間の問題は生じなかった．

実験では SENSEVAL2 の日本語辞書タスク [5] を用いた．決定リスト及び Naive Bayes との判別精度，学習時間，判別時間を比較した．結果，判別精度や判別時間は同等であったが，学習時間は 5 ～ 10 倍高速であった．また決定リスト，Naive Bayes 及び本手法によるアンサンブル学習を試すと，個々の学習手法の判別精度を上回る精度が得られた．

以下，2章で提案する文書検索手法を利用した学習手法を説明する．3章で実験，4章で考察を行い，5章でまとめる．

2 文書検索手法を利用した学習手法

2.1 判別手法

クラス c をもつ訓練事例を集め，その集合を d_c とおく．各事例は素性のリストとして表現される．その各素性を単語とみなしたとき，d_c は単語の集合となり，文書とみなすことができる．分類先のクラスが m 種類あるとし，クラス全体の集合 C を

$$C = \{c_1, c_2, \cdots, c_m\}$$

と表す．すると訓練データから以下のような文書集合 D が作成できる．

$$D = \{d_1, d_2, \cdots, d_m\} \quad (d_i = d_{c_i})$$

テスト事例 t が与えられたとき，再び素性を単語と見なせば，t は単語の集合からなるクエリと見なせる．実際の分類は，テスト事例と最も類似した文書 d_k を求め，クラス c_k を返すことで行える．これは情報検索の問題そのものである．そのためクエリ t と各文書 d_i との距離は，情報検索で使われる手法をそのまま応用できる．

情報検索ではベクトル空間モデルが主流である．ベクトル空間モデルでは，文書とクエリを索引語ベ

クトルで表現し，それらのコサイン尺度などによりそれらの間の類似度を測る．

本手法でもベクトル空間モデルを利用する．ベクトル間の距離は内積により求める．索引語は訓練事例中に現れた全ての単語（すなわち素性）とする．ある索引語に対応する次元の値 d は，通常，その文書内のその索引語の頻度 f であるが，この値の与え方には様々な拡張が存在し，一般には以下の形にまとめられる [7]．

$$d = \frac{l \cdot g}{n} \tag{1}$$

ここで l は局所的重みであり，その文書内のその索引語の出現頻度に基づき計算される重みである．g は大域的重みであり，その文書全体にわたるその索引語の分布を考慮して決定される重みである．そして n は文書正規化係数であり，文書の長さによる影響をなくす目的で導入するものである．

本研究では l として以下の対数化索引語頻度を用いる．

$$l = \log(1 + f)$$

また g や n は考慮せずに，$g = n = 1$ とする．

2.2 索引語リストの作成

本研究では事例を構成する素性が索引語になるために，索引語リストの作成とは事例からどのような素性を抽出するかを意味する．

また本研究では分類問題のタスクを語義判別問題に限定している．そして，語義判別の手がかりとなる属性として以下のものを設定した．

e1	直前の単語
e2	直後の単語
e3	前方の内容語２つまで
e4	後方の内容語２つまで
e5	e3 の分類語彙表の番号
e6	e5 の分類語彙表の番号

例えば，語義判別対象の単語を「出す」として，以下の文を考える（形態素解析され各単語は原型に戻されているとする）．

短い/コメント/を/出す/に/とどまる/た/。

この場合，「出す」の直前，直後の単語は「を」と「に」なので，‘e1=を’，‘e2=に’ となる．次に，「出す」の前方の内容語は「短い」と「コメント」なので，‘e3=短い’，‘e3=コメント’ の２つが作られる．またここでは句読点も内容語に設定しているので，「出す」の後方の内容語は「とどまる」「。」となり，‘e4=とどまる’，‘e4=。’ が作られる．次に「短い」の分類語彙表 [4] の番号を調べると，3.1920_1 である．ここでは分類語彙表の４桁目と５桁目までの数値をとることにした．つまり ‘e3=短い’ に対しては，‘e5=3192’ と ‘e5=31920’ が作られる．「コメント」は分類語彙表には記載されていないので，‘e3=コメント’ に対しては e5 に関する素性は作られない．次は「とどまる」の分類語彙表を調べるはずだが，ここでは平仮名だけで構成される単語の場合，分類語彙表の番号を調べないことにした．これは平仮名だけで構成される単語は多義性が高く，無意味な素性が増えるので，その問題を避けたためである．もしも分類語彙表上で多義になっていた場合には，それぞれの番号に対して並列にすべての素性を作成する．

結果として，上記の例文に対しては以下の８つの素性が得られる．

e1=を，　e2=に，　e3=短い，　e3=コメント，
e4=とどまる，　e4=。，　e5=3192，　e5=31920，

上記の例のようにして，「出す」に対するすべての訓練事例の素性を集める．集められた素性の集合が索引語リストとなる．

2.3 学習と判別

本手法の学習処理は，訓練事例をクラス毎に分類し，クラス毎の事例内の素性を集めるだけである．この処理時間はわずかなものであり，高速な学習が可能である[1]．

[1]本手法は最近傍法と同類の手法であり，学習手法と見なさない立場もある．

一方，実際の判別はテスト事例に対して，各クラスに対応する文書との距離を測るので，クラスの数に比例した処理時間を要する．そのため通常の学習手法よりも判別処理に時間がかかる危険性はある．ただ実際は，距離を測る処理（内積の計算）は高速に行えるし，語義判別に限ればクラスの数も高々10程度なので，実際に問題になることは少ない．

3 実験

SENSEVAL2 の日本語辞書タスク [5] を題材に実験を行う．

日本語辞書タスクは，単純な語義判別問題である．対象単語は名詞 50 単語，動詞 50 単語の計 100 単語である．これら 100 単語は語義の頻度分布のエントロピーを考慮して選定されており，語義判別が容易なものから困難なものまでバランス良く選定されている．ラベル付きの訓練データは 1 単語平均して名詞は 177.4 事例，動詞は 172.7 事例用意されている．またテストデータは各単語に対して 100 問のテストが用意されている．つまり名詞に対しては計 5000 問，動詞に対しても計 5000 問のテストが行える．

この辞書タスクに対して決定リスト，Naive Bayes 及び本手法を用いて解決を図る．使われた素性はみな共通である．決定リストに関しては頻度 1 の素性は間引きした[2]．Naive Bayes はあるクラス c のもとで，素性 f が発生する確率 $P(f|c)$ を求めることにより実現できる．スムージングをどのように行うかが 1 つのポイントだが，ここでは単純に以下の式によりを求めた．

$$P(f|c) = \frac{1 + \sum_{d \in D_c} N(f,d)}{|F| + \sum_{m=1}^{|F|} \sum_{d \in D_c} N(f_m,d)} \quad (2)$$

式 2 の D_c はクラス c を持つ訓練事例の集合を表す．D_c の各要素を d で表す．F は素性全体の集合である．F の各要素を f_m で表す．また，$N(f,d)$ は，訓練事例 d に含まれる素性 f の個数を表す．ここでの設定では，$N(f,d)$ は 0 か 1 の値であり，ほとんどの場合 0 である．

辞書タスクに対する判別精度を表 1 に示す．表中の DL は決定リスト，NB は Naive Bayes を表す．決定リスト，Naive Bayes 及び本手法の精度に大きな差がないことが確認できる．更にそれら 3 つの手法によるアンサンブル学習を行った．最終的な判別は，単純に 3 つの分類器の判別結果の多数決を用いた．この結果も表 1 に示す．名詞に関してはアンサンブル学習が個々の学習手法よりもよい値を出した．また動詞に関しては Naive Bayes の方がアンサンブル学習よりも優れていた．ただし名詞と動詞を総計したものでは，アンサンブル学習が個々の学習手法よりもよい値を出している．

[2]間引きしない場合の実験も行ったが，結果は間引きした方がよかった．

表 1: 精度比較

	DL	NB	本手法	アンサンブル
名詞	0.7618	0.7648	0.7717	0.7748
動詞	0.7769	0.7897	0.7813	0.7849
合計	0.7694	0.7772	0.7765	0.7798

次に決定リスト，Naive Bayes 及び本手法の学習時間と判別時間を計測した．

学習時間に関しては、訓練事例のデータを与えて分類器を得るまでの時間とした．決定リストの場合は決定リストを作成できるまでの時間である．Naive Bayes の場合は各クラスに対する式 2 で示される分布を求めるまでの時間である．そして本手法の場合はクラスに対応する文書の索引語ベクトルを作成するまでの時間である．名詞，動詞合わせて，計 100 単語に対して学習を行い，その合計時間を計測する．この実験を 5 回行い，平均をとった結果を学習時間とした．判別時間に関しては，名詞，動詞合わせて計 10,000 問のテストデータを与えてすべての判別結果を出力するまでの時間とする．この実験を 5 回行い，平均をとった結果を判別時間とした．

3 つの学習手法はすべて perl と shell スクリプトにより実装されているので，開発言語による差はないと考えてよい．用いた計算機は Pentium-4 1.5GHz メモリ 512M バイト，OS は Linux である．計測時間は UNIX の time コマンドの system タイムの出力から得た．

学習時間と判別時間に関する実験結果を表 2 と図 1 に示す．

学習時間に関して本手法は他手法と比べて 5 ～ 10 倍高速であった．また判別時間に関しても他手法と同程度の時間で終了させることができている．

表 2: 学習・判別時間（秒）

	DL	NB	本手法
学習時間	3.070	5.638	0.558
判別時間	0.316	0.608	0.410

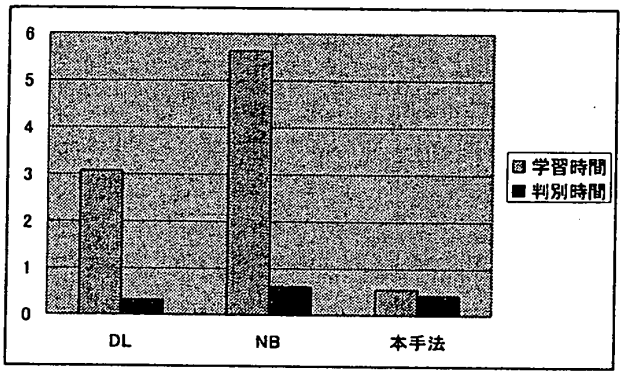

図 1: 学習・判別時間（秒）

4 考察

本手法は k-最近傍法と類似点がある．k-最近傍法は，テスト事例 x_q が与えられたときに各訓練事例との距離を測り，距離が短い順に k 個の訓練事例 x_i $(i = 1, 2, \cdots, k)$ を取りだし，それらのクラスの多数決によりクラスを決定する手法である．

k-最近傍法は距離による重み w_i を乗じることで、以下のように拡張できる [3]．

$$\hat{f}(x_q) \leftarrow arg \max_{c \in C} \sum_{i=1}^{k} w_i \delta(c, f(x_i)) \qquad (3)$$

ここで f は訓練事例に対してそのクラスを返す関数である．$\hat{f}$ は任意の事例に対して判別先のクラスを返す関数であり，目的の分類器である．$\delta(c, f(x_i))$ は x_i のクラス $f(x_i)$ が，クラス c と等しいときに 1，そうでないとき 0 を返す関数である．w_i は，通常，x_q と x_i との距離 $d(x_q, x_i)$ の 2 乗の逆数で与えられる．

$$w_i = \frac{1}{d(x_q, x_i)^2}$$

式 3 において．k を訓練事例の全体の数 N とした場合を考えてみる[3]．本来，重み w_i は x_q と x_i 間の距離が短いほど大きな値をとるように設定すればよいので，その内積で定義することにする．

$$w_i = (x_q, x_i)$$

いま，クラスが c であるような訓練事例が m 個あり，それらを x_i $(i = 1, 2, \cdots, m)$ で表す．このとき

$$\sum_{i=1}^{N} w_i \delta(c, f(x_i)) = \sum_{i=1}^{m} (x_q, x_i) = (x_q, \sum_{i=1}^{m} x_i)$$

が成立する．$\sum_{i=1}^{m} x_i$ は，x_i が素性ベクトルであることを考えれば，本手法によってクラス c の事例の素性を集めて，各素性の頻度により新たに作られたベクトルに対応する．つまり式 3 による判別は，本手法において式 1 における l, g, n の値を以下のように設定した場合に相当する．

$$l = f, \quad g = n = 1$$

式 1 において $l = f$, $g = n = 1$ とおいた場合の辞書タスクに対する判別精度は以下の通りであった．局所的重みを考慮している本手法の効果が確認できる．

表 3: 精度比較

	$l = f$ (Shepard)	$l = \log(1 + f)$ (本手法)
名詞	0.7695	0.7717
動詞	0.7582	0.7813
合計	0.7638	0.7765

ここでは本手法を語義判別問題に適用したが，他の分類問題についても適用できる可能性がある．他の分類問題に対してもある程度の精度が出せるのかどうかの調査は今後の課題である．ここでは試しに MLC++[2] のデータとして配布されている waveform-40 という分類問題に本手法を適用してみた[4]．この分類問題は分類先クラスが 3 つであり，その分布は一様である．また 40 個の属性（うち 19 個はノイズ）が用意され，それらは実数値をとる．訓練事例数は 300，テスト事例数は 4700 である．

[3] f が実数値関数の場合，この判別手法は Shepard's method と呼ばれる [3].

[4] 属性値が実数値なので語義判別問題とはタイプが異る，最近傍法での精度が極端に悪い，という 2 つの観点から選択した．

Bayes 分類器で 86%，決定木で 72%，最近傍法で 38%の精度が報告されている．本手法を試した結果は 68% の精度が得られた．極端に悪いというレベルではなかった．本手法は最近傍法の改良に位置づけられる手法であるために，ある程度の頑健性は備えていると考えられる．

5 おわりに

ここでは情報検索手法を利用した分類問題の解決手法を提案した．k-最近傍法と類似の手法であり，学習時間をほとんど要しないという特徴がある．そのためアンサンブル学習の 1 つの学習器としての利用が可能である．

SENSEVAL2 の辞書タスクによる実験では，本手法は決定リストや Naive Bayes と同程度の精度を出し，判別時間も同程度であった．一方，学習時間に関しては 5 ～ 10 倍高速であった．また決定リスト，Naive Bayes 及び本手法によるアンサンブル学習により，個々の学習手法の判別精度を上回る精度が得られた．

今後の課題としては，本手法を様々な問題に適用し，その特徴を調べることである．特に辞書タスク以外の語義判別問題，あるいは他の分類問題へ適用し，本手法がどのような問題に対して有効であるのかを調査したい．

参考文献

[1] Ethem Alpaydm. Techniques for combining multiple learners. In *Engineering of Intelligent Systems*, Vol. 2, pp. 6–12, 1998.

[2] Ron Kohavi, Dan Sommerfield, and James Dougherty. Data mining using MLC++: A machine learning library in C++. Vol. 6, No. 4, pp. 537–566, 1997.

[3] Tom Mitchell. *Machine Learning*. McGraw-Hill Companies, 1997.

[4] 国立国語研究所. 分類語彙表. 秀英出版, 1994.

[5] 黒橋禎夫, 白井清昭. SENSEVAL-2 日本語タスク. 電子情報通信学会言語とコミュニケーション研究会, NLC-36～48, pp. 1–8, 2001.

[6] 上田修功, 中野良平. アンサンブル学習における汎化誤差解析. 電子情報通信学会論文誌, No. 9, pp. 2512–2521, 1997.

[7] 北研二, 津田和彦, 獅々堀正幹. 情報検索アルゴリズム. 共立出版, 2001.