ファジィクラスタリングを用いた語義判別規則の教師なし学習

○ 高橋篤史
茨城大学大学院理工学研究科
システム工学専攻

新納浩幸
茨城大学大学工学部システム工学科

1 はじめに

本論文ではファジィクラスタリングを用いて語義判別規則の教師なし学習を試みる．

自然言語処理では個々の問題を分類問題として定式化し，帰納学習の手法を利用して，その問題を解決するというアプローチが大きな成功をおさめている．しかしこのアプローチには帰納学習で必要とされる訓練データを用意しなければならないという大きな問題がある．

近年，この問題に対する1つの対処法として少量のラベル付き訓練データから得られる分類規則の精度を，大量のラベルなし訓練データによって高めてゆく教師なし学習が提案されている．ここではクラスター分析の一手法であるファジィクラスタリングを利用した教師なし学習を試みる．ファジィクラスタリングは，個体がクラスに帰属する度合いに曖昧さを認めるという考えに基づいている．曖昧さの表現はファジィ理論による要素への帰属度によって表される．ここでは，まずラベル付きの各訓練事例をクラスのプロトタイプに設定し，次にファジィクラスタリングを用いてラベル付き及びラベルなしの全事例の各クラスへの帰属度を求める．この処理により各クラスのプロトタイプがより適切な位置へ移動する．実際の識別は，移動された後のクラスのプロトタイプの集合を訓練事例として，k-最近傍法により行う．最初に用意されたラベル付き訓練事例の集合と k-最近傍法を用いて識別を行った場合が，通常の学習に対応し，ファジィクラスタリングを行った後の訓練事例の集合と k-最近傍法を用いて識別を行った場合が，教師なし学習に対応する．

実験では SENSEVAL2 の日本語辞書タスク [5] を題材にした．通常の k-最近傍法による正解率は名詞 76.83%，動詞 77.79%であった．一方，本手法の正解率は名詞 76.07%，動詞 77.83%であった．全体的にみると精度はほとんど向上していなかったが，個別に見ると精度が向上している単語も多かった．有効な利用方法などを考察する．

2 ファジィクラスタリングによる多義語の曖昧性解消

2.1 k-最近傍法による識別

最近傍法とは分類問題に対する識別手法の1つである．そこでは入力事例 y と最も距離が近い訓練事例 x を選び，x のクラスを出力する．

k-最近傍法とは最近傍法を拡張した手法であり，y に最も近いものから順に k 個の訓練事例を選び，これら k 個の事例のクラスの多数決によってクラスを識別する [1]．

2.2 ファジィクラスタリングによる教師なし学習

まずラベル付きの各訓練事例を1つのクラスに対応させる．その事例自身がクラスタであり，しかもクラスのプロトタイプにもなっている．次にラベル付き及びラベルなしの全事例をそれらクラスタにクラスタリングさせる[1]．このクラスタリングの手法としてファジィクラスタリングを用いる．

クラスタリングによって最初に設定してあるプロトタイプがより適切な位置に移動する．移動した先のプロトタイプをラベル付きの訓練事例と見て，k-最近傍法によって識別を行う．これは一種の教師なし学習である．

例えば，図1(a) では事例 x とプロトタイプ c_1 との距離と，x とプロトタイプ c_2 との距離が等しいので，x のクラスが C_1 あるいは C_2 かは判定できない．一方，クラスタリングを行うことで，図1(b) のようにプロトタイプ c_1 や c_2 がより適切な位置に移動することで x のクラスが判定できる．

[1]ラベル付きの事例はクラスが確定しているが，クラスタリングではラベル付きの事例も同時にクラスタリングする必要がある．

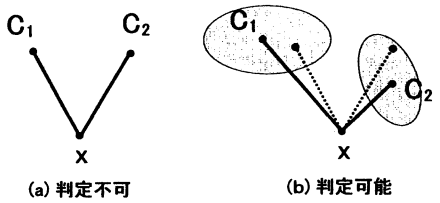

図 1: クラスタリングによるプロトタイプの移動

ここではファジィクラスタリングの処理を説明する [3]. まずクラス（具体的にはラベル付きの事例）が m 種類あるとして，クラスの集合は以下で表せる.

$$C = \{C_1, C_2, \cdots, C_m\}$$

そしてクラスタリングさせる個体（具体的にはラベル付き及びラベルなしの全事例）が n 個あるとし，各個体は p 次元ユークリッド空間上の点で表せるとする. k 番目の個体 x_k は以下のような列ベクトルで表せる.

$$x_k = (x_k^1, \ldots, x_k^p)^T$$

個体 x_k がクラス C_i に帰属する程度を u_{ik} で表す. 通常のクラスタリングはこの値が 0 か 1 となるが，ファジィクラスタリングの場合，0 から 1 の実数値となる. そして $m \times n$ の行列 $U = (u_{ik})$ を定義しておく. またクラス C_i についてプロトタイプを 1 つ選び，そのプロトタイプを行ベクトル

$$v_i = (v_i^1, \ldots, v_i^p)^T$$

で表す. また $p \times m$ の行列 $V = (v_i^k)$ を定義しておく.

ファジィクラスタリングでは，ある目的関数の値を減少させるように，V と U を更新してゆく. 目的関数は様々なものが提案されているが，ここでは代表的な以下の関数を用いる.

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{m}(u_{ik})^r \|x_k - v_i\|^2 \qquad (1)$$

式 1 の r は $r > 1$ を満たすように選ばれるパラメータであり，ここでは $r = 2$ に設定した.

次にこの目的関数を減少させるように V と U を更新してゆくアルゴリズム **FCM** を以下に示す. これはファジィc-平均法と呼ばれている.

FCM(Fuzzy c-Means)

FCM1. $\bar{V}$ の初期値を定める.

FCM2. $\bar{V}$ を固定して，

$$\min_{U \in M_f} J(U, \bar{V})$$

を解き，最適解を $\bar{U}$ とする.

FCM3. $\bar{U}$ を固定して，

$$\min_{V} J(\bar{U}, V)$$

を解き，最適解を $\bar{V}$ とする.

FCM4. 解 $(\bar{U}, \bar{V})$ が収束すれば終了. そうでなければ **FCM2** へ.

FCM2 における最適解は次の式で与えられる.

すべての $v_i (i = 1, \ldots, m)$ に対して $x_k \neq v_i$ であるような x_k については，

$$u_{ik} = \left[\sum_{j=1}^{m}\left(\frac{\|x_k - \bar{v}_i\|^2}{\|x_k - \bar{v}_j\|^2}\right)^{\frac{1}{r-1}}\right]^{-1}, \quad for\ x_k \neq v_i, i = 1, \ldots, m \qquad (2)$$

また，ある v_i について，$x_k = v_i$ となる x_k については，

$$u_{ik} = 1; \quad u_{jk} = 0\ (j \neq i). \qquad (3)$$

FCM3 における最適解はは次の式で与えられる.

$$v_i = \frac{\sum_{k=1}^{n}(\bar{u}_{ik})^r x_k}{\sum_{k=1}^{n}(\bar{u}_{ik})^r} \qquad (4)$$

FCM の収束条件については，様々なバリエーションがあるが，ここでは最も単純に最大繰り返しの回数を設定することにした. 実験ではこの最大繰り返しの回数は 5 とした.

また **FCM** を本手法で用いる際にラベル付き事例の扱いには注意が必要である. 1 回目のループの **FCM2** のステップでは，ラベル付き事例はクラスのプロトタイプと一致しているので，式 3 が用いられる. 2 回目以降のループの **FCM2** のステップでは，一般に，ラベル付きの事例はプロトタイプと一致していない. しかしこの場合でもラベル付きの事例はクラスが確定しているので，式 2 ではなく式 3 を用いる.

2.3 利用した素性集合

ファジィクラスタリングを行うためには，各事例を p 次のベクトルで表現すればよい. 一般に分類問題の

事例は素性の集合で表されるので，各素性を次元に割り当て，その素性が存在する場合にその次元の値を 1 に設定し，存在しない場合に 0 に設定することで，各事例が p 次のベクトルで表現できる．

ここではで利用した素性について述べる．

まず語義の曖昧性解消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ている．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，

3 実験

本手法の有効性を確認するために，SENSEVAL2 の日本語辞書タスク [5] で課題とされた名詞 50 単語，動詞 50 単語の多義語の曖昧性解消を試みた．

SENSEVAL2 の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．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．またラベルなし訓練データは RWC テキストデータベース第 2 版に納められた毎日新聞 95 年度版の 1 年分の記事を利用して，1 単語平均して名詞は 7585.5 事例，6571.9 事例を収集した．

実験結果を表 1 に示す．表中の k-NN の列はラベル付きの訓練事例のみを用いた k-最近傍法 ($k = 5$) の正解率 (%) を示している．また表中の Fuzzy の列は本手法の正解率を (%) を示している．平均をとると表 2 のようにまとまる．

4 考察

教師なし学習を試みたが，全体の精度はほぼ変化がなかったと言える．これは本手法の効果が低いことを意味してはいない．それは個々の単語で見ると，正解率が大きく向上するものや大きく低下する単語が存在するからである．例えば，ukeru(+13%)，susumu(+10%)，matsu(+8%)，doujitsu(+6%) などは大きく正解率が向上しているが，ima(-13%)，shimin(-11%)，kodomo(-10%)，tsukau(-9.75%) などは大きく正解率が低下している．つまり本手法はある単語については効果があるが，ある単語については逆効果になっている．

ラベル付き訓練データの他にラベルなし訓練データを用いる教師なし学習は，ラベル付き訓練データのみを用いる通常の学習よりも，精度が低くなる場合もある．これは同じデータで教師なし学習を行った研究でも確認されている [2]．この場合の対処法としては 2 つあると考える．1 つは論文 [2] のように効果がある問題とない問題を予め予測して，適用の有無や程度を調整するアプローチである．もう 1 つは副作用が少ない，つまり頑健性が高い教師なし学習を用いることである．そして本手法は比較的頑健性が高いと思われる．

例えば論文 [2] では教師なし学習として EM アルゴリズムを用いた手法を試みているが，何の対処も行わ

表 1: 実験結果

名詞	k-NN	Fuzzy	動詞	k-NN	Fuzzy
aida	79	71	ataeru	68	68
atama	65	60	iu	94	93
ippan	88.33	90.67	ukeru	49	62
ippou	87	86	uttaeru	83	88
ima	85	72	umareru	68	70
imi	55	59	egaku	57	53
utagai	100	100	omou	90	91
otoko	92	92	kau	85	85
kaihatsu	64	64	kakaru	60	64
kaku_n	72	73	kaku_v	75	74
kankei	84	86	kawaru	92	92
kimochi	62	67	kangaeru	99	99
kiroku	63	63	kiku	65	63
gijutsu	96	95	kimaru	96	96
genzai	98	98	kimeru	92	91
koushou	100	100	kuru	85	82
kokunai	53	53	kuwaeru	89	89
kotoba	51	50	koeru	78	84
kodomo	70	60	shiru	97	97
gogo	88.5	83.5	susumu	42	52
shijo	71	67	susumeru	97	96
shimin	63	52	dasu	30	31
shakai	79	80	chigau	100	100
shonen	92	92	tsukau	95.5	85.75
jikan	54	53	tsukuru	66	64
jigyou	68	72	tsutaeru	75	76
jidai	72	77	dekiru	81	78
jibun	100	96	deru	54	52
joho	74	71	tou	71	68
sugata	59	61	toru	32	33
seishin	66	65	nerau	99	99
taishou	96	96	nokosu	79	75
daihyou	85.5	89.5	noru	57	61
chikaku	81	82	hairu	37	39
chihou	74	66	hakaru	93	93
chushin	98	98	hanasu	100	100
te	46	47	hiraku	85	87
teido	99	99	fukumu	99	97
denwa	85	85	matsu	44	52
doujitsu	65	71	matomeru	80	72
hana	99	99	mamoru	79	73.5
hantai	98	98	miseru	97	96
baai	90	86	mitomeru	89	89
mae	87	87	miru	81	78
minkan	100	100	mukaeru	89	89
musume	86	84	motsu	55	52
mune	63	67	motomeru	87	87
me	13	13	yomu	88	88
mono	30	32	yoru	96	97
mondai	95	95	wakaru	90	90
平均	76.83	76.07	平均	77.79	77.83

ずにそのまま適用すると，名詞は 77.54% から 72.82% に正解率が下がり，動詞は 78.44 % から 78.70 % に正解率が上がる．全体としては 77.99% から 75.76%，つまり 2.23% の正解率の低下があるが，本手法の正解率の低下は 0.34% である．これは本手法では極端に正解率が悪くなることがないことを示している．

また本手法により得られたプロトタイプは新たな事例として扱えるという長所も持つ．このため，様々な展開が考えられる．例えば，得られた事例から別の学習手法を用いて分類器を作成することも可能であろう．

本手法の欠点としては計算時間が膨大になる点がある．アルゴリズム **FCM** はプロトタイプの数（つまりラベル付き訓練データの数）が多いと計算時間が多大にかかる．本実験では Pentium-4 1.5GHz のマシンで，1 回の繰り返しに 1〜2 時間はかかった．計算を効率よく行う工夫は今後の課題である．

またベースとした k-最近傍法自体あまり識別精度が高くなかったので，この点も改良しなくてはならない．

表 2: 正解率の比較 (%)

	k-NN	本手法
名詞	76.83	76.07
動詞	77.79	77.87
合計	77.31	76.97

とくに事例をベクトルで表現する際に，素性のあるなしで 1 か 0 の値をつけるのは適切ではない．素性の重みを考慮すべきである．また k の値をここでは 5 にしたが，単語によって最適な k の値は異なるであろう．これらの点からの改良も今後の課題である．

5 おわりに

本論文ではファジィクラスタリングを用いて語義判別規則の教師なし学習を試みた．ラベル付きの事例をクラスに設定し，ラベル付き及びラベルなしの全事例を使ってファジィクラスタリングすることで，クラスのプロトタイプをより適切な位置に移動させる．移動後のプロトタイプをラベル付きの事例として扱い，k-最近傍法により識別を行なう．実験では SENSEVAL2 の日本語辞書タスクを用いた．平均的にはラベルなし事例を用いた効果は現れなかったが，各単語でみると効果がある単語も多く，有効利用できる可能性もある．今後は個別の単語ごとにラベルなし事例の利用を調整する方法を考えたい．また識別のベースとした k-最近傍法も改良したい．

参考文献

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

[2] 新納浩幸, 佐々木稔. EM アルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師なし学習. 情報処理学会自然言語処理研究会, NL-151-8, 2002.

[3] 宮本定明. クラスター分析入門. 森北出版, 1999.

[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.