自然言語処理　144－13
(2001. 7. 16)

ベイズ推定による決定リストのルール信頼度推定法

鶴岡 慶雅　近山 隆

東京大学工学系研究科　東京大学新領域創成科学研究科

〒113–8656 東京都文京区本郷 7-3-1
東京大学工学系研究科電子情報工学専攻近山研究室
03-5613-7069
tsuruoka@logos.t.u-tokyo.ac.jp, chikayama@klic.org

あらまし　統計的クラス分類器としての決定リストは，近年自然言語処理における様々な分野でその有効性を示している．決定リストを構成する上で，もっとも重要な問題の一つは，ルールの信頼度の算出法である．しかし決定リストを用いた多くの研究では，信頼度の算出法についてそれほど注意が払われていない．そこで本論文では，ベイズ学習法を利用してルールの信頼度を算出する手法を示す．さらに，証拠の種類ごとに異なる事前分布を利用することで，より正確な信頼度の推定が可能になり，決定リストの性能が向上することを示す．本論文では，本手法の有効性を確かめるために，疑似単語の判定問題に関して実験を行った．その結果，事前分布を用いることにより判定精度の向上と，判定の信頼度自身の正確さの向上という，2つの点で効果が見られた．

キーワード　決定リスト，ベイズ推定，事前分布

Estimating reliability of the rules in decision lists using Bayesian learning

Yoshimasa Tsuruoka and Takashi Chikayama

School of Engineering, The University of Tokyo
School of Frontier Sciences, The University of Tokyo

7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
03-5613-7069
tsuruoka@logos.t.u-tokyo.ac.jp, chikayama@klic.org

Abstract　The decision list algorithm is one of the most successful algorithms for classification problems in natural language processing. We propose a method based on Bayesian learning to calculate the reliability of contextual evidences in decision lists. The method also gives well-founded smoothing and better use of prior information of each type of contextual evidence. We evaluate these proposed methods on Japanese word sense disambiguation problems. The results show improved accuracy close to one expected from the Bayesian theory.

key words　Decision list, Bayesian learning, Prior distribution

1 はじめに

決定リストとは統計的なクラス分類器である．自然言語処理の多くは，クラス分類問題として捉えることが可能であり，近年，様々な自然言語処理において，決定リストによる手法の有効性が示されている [3][4]．特に，語義曖昧性解消問題に対しては，語義曖昧性解消システムの性能を競う大会である Senseval-1 において，決定リストを階層的に拡張した手法が最も良い成績をあげている [6]．

クラス分類器としては，分類精度の点だけでいえば，最近ではサポートベクタマシン（SVM）やアダブースト（AdaBoost）といった手法が，その性能の高さから注目を集めている．しかし，決定リストによる手法では，作成された分類器が if-then 形式のルールの並びであるために，人間が容易に理解可能であるというメリットがある．それにより，自動的に学習された決定リストに人間の手を入れることで，性能を向上させることができるとの報告もある [1]．

決定リストを作成する上で最も重要な問題は，ルールの信頼度の算出法である．事例の数が多いルールに関しては，ベルヌーイ試行における最尤推定法により，出現頻度の比として計算してもそれほど問題にはならない．しかし，事例の数が少ない場合，単に出現頻度の比を用いると，その値は非常に不正確な値となってしまう．従来，この問題に対しては，事例の数が少ないルールは使用しない方法や，ルールの信頼度の算出の式に適当な数値を足すことによるスムージングなどの対処がなされていた．しかし，ルールを間引く前者の手法では，重要なルールを取りこぼしてしまう危険があり，後者の，計算式に適当な数値を足してスムージングを行なう手法では，加算する値の設定に関して理論的な指針に欠けるという問題がある．

そこで本論文では，ルールの確率値の推定にベイズ推定の手法を用いる．さらに，ベイズ推定の枠組における事前情報として，証拠のタイプごとに異なる事前分布を用いることによって，より正確な推定が可能になり，決定リストの性能を向上させることができることを示す．

本論文の構成は以下の通りである．2 章で決定リストによるクラス分類の手法を説明する．3 章で，ベイズ推定による確率値の算出法を示す．4 章で，事前分布を利用する方法を示す．5 章で，疑似単語を用いた実験結果を示す．

2 決定リスト

決定リストとは，クラス分類のためのルールを，その信頼度の高い順に並べたものである．それぞれのルールは，「もし（証拠 E_i）ならば，クラスは C_j である」という形式をしている．証拠というのは，判定の手がかりとなる事例の特徴である．

例として，英語の多義語 *plant*（A: 植物,B: 工場）に関して Yarowsky が行なった実験での決定リストを表 1 に示す [5]．最上位のルールは，「右隣に life という単語があったら，語義は A」という意味，4 番目のルールは，「距離 2〜10 単語以内に manufacturing という単語があったら，語義は B」という意味である．

実際にクラスの分類を行なう際には，その事例に対して適用可能なルールのうち，最も上位のルールを用いて分類が行なわれる．例えば入力文が，

... divide life into *plant* and animal kingdom ...

であるとすると，適用可能なルールのうち最上位なのは 3 番目のルールであるから，*plant* の語義は A だと判定されることになる．

決定リストは，事例とその正解ラベルを含む訓練コーパスから作成される．決定リストの作成において，もっとも重要な問題は，それぞれのルールの信頼度の計算法である．文献 [5] では，次の式に従って信頼度を計算している．

$$(\text{信頼度}) = \log\left(\frac{P(C_A|E_i)}{P(C_B|E_i)}\right) \tag{1}$$

すなわち，証拠 E_i のもとでクラス（語義）が A である確率と，同じ証拠 E_i のもとでクラスが B である確率との比の対数をとったものである．

3 ベイズ推定によるルールの確率値の推定

従来の決定リストを用いた自然言語処理の研究では，ルールの信頼度の算出法として，式 (1)，あるいは，式 (1) をクラスが 3 つ以上の場合でも適用できるように変形した次の式，

$$(\text{信頼度}) = \log\left(\frac{P(C_A|E_i)}{1 - P(C_A|E_i)}\right) \tag{2}$$

が用いられることが多い [5][6][3]．また，対数をとらずに，

表 1: 決定リストの例

信頼度	証拠	語義
8.10	*plant* **life**	A
7.58	**manufacturing** *plant*	B
7.39	**life** (within ±2–10words)	A
7.20	**manufacturing** (within ±2–10words)	B
6.27	**animal** (within ±2–10words)	A
4.70	**equipment** (within ±2–10words)	B
4.39	**employee** (within ±2–10words)	B
:	:	:

$$(\text{信頼度}) = P(C_A|E_i) \tag{3}$$

とする場合もある．

ここで，式（3）と式（2）を見比べてみると，式（2）は，式（3）に関して単調増加であり，決定リストでは信頼度の大小関係しか問題にならないのだから，(後述するスムージングの問題を考慮しなければ）式（2）を用いた場合と，式（3）を用いた場合では，結果的に作成される決定リストは等価になる．また，一般にクラス分類器の目標は，分類の正解率を最大にすることであるから，ルールの信頼度としては，そのルールが正解する確率である式（3）を用いるのが自然である．そこで，そこで本論文では，ルールの信頼度として式（3）を用いることにする．

式（3）の値は，訓練事例が多ければ，ベルヌーイ試行における最尤推定により，次のように計算することができる．

$$(\text{信頼度}) = P(C_A|E_i) = \frac{f(C_A, E_i)}{f(E_i)} \tag{4}$$

ただし，$f(C_A, E_i)$ は，クラス A に属するターゲットと証拠 E_i が同時に出現した回数．$f(E_i)$ は，証拠 E_i の出現回数である．ところが，通常は出現回数が少ない証拠も多い．例えば，

$$f(C_A, E_i) = 1, f(E_i) = 1 \tag{5}$$

の場合，信頼度は $1/1 = 1$ と計算されるが，たった一つの事例しかないのに，その信頼度が 100%になるというのは直感的にもおかしいし，実用上も，このように計算して作成した決定リストの分類性能は非常に低くなってしまう．このように，出現回数の少ない事例において，そのままでは統計的に信頼性のある確率値が算出できないことをスパースネスの問題という．

そこで本論文では，ベイズ推定の手法を用いてこの問題の解決を試みる．

まず，ルールの確率を確率変数 $\tilde{\theta}$ として考え，ルールの確率を求める問題を，確率変数 $\tilde{\theta}$ の期待値を求める問題と考えることにする．

いま，訓練コーパスにおいて，確率を求めたいルールに関する事例が n 個あり，そのうちの k 個において，そのルールが正解だったとする．この事象を A とする．

ベイズの定理に従えば，事象 A が起こったという条件のもとでの $\tilde{\theta}$ の事後密度は，

$$P(\theta|A) = \frac{P(\theta)P(A|\theta)}{P(A)} \tag{6}$$

$$= \frac{P(\theta)P(A|\theta)}{\int_0^1 P(\theta)P(A|\theta)d\theta} \tag{7}$$

で与えられる．ここで，事象 A はベルヌーイ試行と考えられるから，その確率は二項分布により次のように与えられる．

$$P(A|\theta) = {}_nC_k\theta^k(1-\theta)^{n-k} \tag{8}$$

これを式（7）に代入して，

$$P(\theta|A) = \frac{P(\theta){}_nC_k\theta^k(1-\theta)^{n-k}}{\int_0^1 P(\theta){}_nC_k\theta^k(1-\theta)^{n-k}d\theta} \tag{9}$$

$$= \frac{P(\theta)\theta^k(1-\theta)^{n-k}}{\int_0^1 P(\theta)\theta^k(1-\theta)^{n-k}d\theta} \tag{10}$$

を得る．ここで，事前確率密度 $P(\theta)$ をどのように設定するのか，という問題が浮上する．以下，事前分布の設定について，2 つの方法を提案する．最初に，一様分布とする方法，次にルールの確率値の分布をベータ分布で近似する手法である．

3.1 一様分布

まず，あるルールの確率に関して，事前知識が全くないものと考えると，すべての確率値の事前確率について同じ値とするのが自然である．すなわち，事前確率密度は一様分布とする．

$$P(\theta) = 1 \tag{11}$$

そうすると，事後分布は次のようになる．

$P(\theta|A)$

$$= \frac{\theta^k(1-\theta)^{n-k}}{\int_0^1 \theta^k(1-\theta)^{n-k}d\theta} \tag{12}$$

$$= \frac{\theta^{(k+1)-1}(1-\theta)^{(n+2)-(k+1)-1}}{\int_0^1 \theta^{(k+1)-1}(1-\theta)^{(n+2)-(k+1)-1}d\theta} \tag{13}$$

この確率分布は，ベータ分布と呼ばれ，期待値は次式で与えられる [7]．

$$E[\theta] = \frac{k+1}{n+2} \tag{14}$$

いま，k と n は，それぞれ，$f(C_A, E_i)$ と $f(E_i)$ に対応しているのだから，

$$(\text{信頼度}) = P(C_A|E_i) = \frac{f(C_A, E_i)+1}{f(E_i)+2} \tag{15}$$

となる．結論は非常にシンプルである．すなわち，頻度 $f(C_A, E_i)$ と $f(E_i)$ をそのまま用いる代わりに，$f(C_A, E_i)+1$ と $f(E_i)+2$ を用いればよい，ということである．

表 2: 実験に用いた疑似単語

品詞	疑似単語
名詞	政策（48.5%）／テレビ（51.5%）
	大統領（42.9%）／首相（57.1%）
	仕事（26.2%）／言葉（26.9%）／資金（26.5%）／文化（20.4%）
動詞	持つ（63.2%）／含む（36.8%）
	考える（36.2%）／見る（32.2%）／目指す（31.6%）
	入る（32.2%）／示す（23.9%）／開く（22.5%）／進める（21.4%）
形容詞	近い（52.0%）／難しい（48.0%）
	新しい（30.2%）／高い（33.9%）／強い（35.8%）
	若い（27.0%）／厳しい（25.6%）／大きい（22.8%）／よい（24.5%）

（括弧内はテストコーパス中での出現割合）

3.2 実験

3.2.1 実験用コーパスと疑似単語

最初に，事前分布を一様分布とした場合の基本的な実験を行なう．本論文では，手法の評価を，疑似単語の判定を決定リストで行なう実験によって行なう．

疑似単語とは，複数の異なる単語を判定システムの側からは同一の単語にしか見えないようにし，文脈からどの単語であるのかを判定させる手法である．例えば，「銀行」という単語と，「土手」という単語を用いて疑似単語を作ったとすると，判定システムからは，入力文は例えば，

... お金をおろしに＊＊へ行く途中 ...

のように見える．＊＊の部分が疑似単語である．そして，文脈から，「銀行」であるのか「土手」であるのかを判定させるというわけである．これは，文脈から多義語の語義の判定を行う多義性解消の問題とかなり似た問題になる．

実験に用いる疑似単語に関しては，訓練事例を大量に得るため，出現回数の多い単語の中から選んだ．また，ベースラインとしての正解率（単純に最も出現頻度の高い単語を選ぶ方法の正解率）が高くならないように，一つの疑似単語を構成する各々の単語の出現頻度がほぼ等しくなるようにして構成した．表 2 に，実験に用いた疑似単語と，それを構成する単語のテストコーパス中における出現割合を示す．

コーパスとしては，「CD-毎日新聞 97 年版」を JUMAN version 3.6 [8] で形態素解析したものを用いた．

事例の数に関しては，各々の疑似単語について，1024 の訓練事例，1000 のテスト事例を重なりがないようにコーパスからランダムに抽出して，トレーニングとテストを行なった．

3.2.2 分類に用いる証拠

本論文では，クラス判定のための文脈情報として，以下に示す証拠を用いる．

- Adjacent

  ターゲットの左隣に出現する単語（A_{-1} と略記）．

  ターゲットの右隣に出現する単語（A_{+1} と略記）．

- Window

  ターゲットから，距離 10 単語以内に出現する単語（W と略記）．

- Pair

  ターゲットの左隣にある単語対（P_{-1} と略記）．

  ターゲットを挟む単語対（P_0 と略記）．

  ターゲットの右隣にある単語対（P_{+1} と略記）．

例えば，ターゲット「環境」を含む次の文

「自然 環境 を 大切 に する」

に対しては，以下の証拠が生成されることになる．

（A_{-1} 自然），（A_{+1} を），（W 自然），（W を），（W 大切），（W に），（W する），（P_0 自然 を），（P_{+1} を 大切）．

3.3 結果

表 3 にその結果を示す．表中の数字は正解率である．また，正解率の隣の括弧内の数字は，判定に用いたルールの信頼度の平均（以後，期待正解率と呼ぶ）である．adjacent, window, pair とあるのは，それぞれの証拠を単独で用いた場合の正解率であり，all とあるのは，全ての証拠を用いた場合の正解率である．

一見してわかるように，ほぼ全ての疑似単語において，all の正解率が最も高くなっている．すなわち，決定リス

表 3: 決定リストによる疑似単語の判定の正解率

品詞	疑似単語	adjacent	window	pair	all
名詞	政策／テレビ	85.9% (80.0%)	92.1% (89.6%)	81.7% (72.0%)	92.0% (90.5%)
	大統領／首相	83.3% (80.4%)	86.3% (87.2%)	82.2% (77.3%)	88.6% (88.6%)
	仕事／言葉／資金／文化	62.5% (62.8%)	66.8% (77.3%)	59.8% (59.6%)	71.6% (81.2%)
動詞	持つ／含む	81.6% (77.0%)	82.4% (85.0%)	87.0% (80.3%)	86.6% (88.9%)
	考える／見る／目指す	70.1% (63.9%)	62.8% (75.4%)	70.7% (67.5%)	70.5% (78.9%)
	入る／示す／開く／進める	61.7% (64.6%)	60.3% (75.6%)	75.6% (71.3%)	76.3% (83.1%)
形容詞	近い／難しい	89.5% (86.4%)	77.4% (82.2%)	85.2% (74.1%)	88.6% (90.6%)
	新しい／高い／強い	61.8% (61.8%)	66.4% (77.4%)	63.2% (60.6%)	68.7% (78.8%)
	若い／厳しい／大きい／よい	68.6% (65.4%)	64.0% (75.0%)	68.3% (61.2%)	72.0% (79.8%)
平均		73.9% (71.5%)	73.2% (80.5%)	74.9% (69.3%)	79.4% (84.5%)

（括弧内は判定に用いたルールの確率値の平均値）

トによって，複数の異なる種類の証拠が適切に統合されていることを示している．

本手法は，ルールの信頼度として式 (15) を用いることにより，判定の信頼度を確率として算出している．本手法による確率値の推定が適切であれば，期待正解率は実際の正解率と近い値になるはずである．そこで，表 3 の，実際の正解率と，期待正解率とを比較してみると，多少のずれはあるものの，ほぼ実際の正解率に近い値が得られている．このことから，ベイズ推定によるルールの確率値の推定が，ある程度有効であるといえる．しかし細かく見ると，window では，期待正解率が実際の正解率よりも高く，逆に pair では低い傾向がある．

4 事前分布の利用による確率値の正確な推定

前章では，ベイズ学習において事前情報が全くないものとし，事前分布を一様分布として計算を行なった．しかし，表 3 からわかるように，実際の正解率と，期待正解率との間には，まだある程度の誤差が存在する．その大きな原因は，事前分布の設定にあると考えられる．そこで，本章では，事前分布を一様分布ではなく，ルールの確率値の性格を反映した分布にすることにより，より正確な推定を試みる．

まず，ルールの確率値の分布がどのような性格を持っているのかを知るために，実際のルールの確率値の分布の例を図 1 に示す．これは，疑似単語「入る／示す／開く／進める」において，証拠のタイプが window のルールの確率値の相対頻度を示したものである．ただし，各々のルールの確率値は，事前分布を一様分布としたベイズ推定により算出している．また，事例の数が少ない証拠は，確率の推定値の誤差が大きいため，事例の数が 5 個未満のルールは除外している．この図からわかるように，ルールの確率値は，実際には一様分布のようなフラット

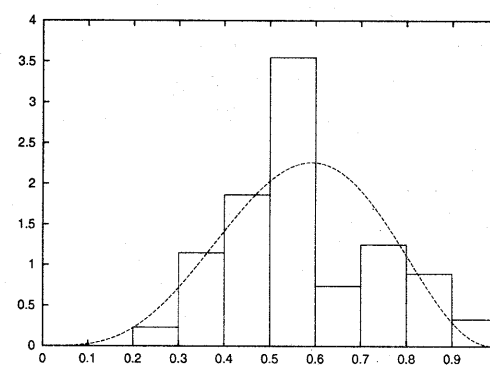

図 1: ルールの確率値の分布

名分布ではなく，ある程度，集中した分布になっていることがわかる．

この図では，事例の数が 5 個未満のルールは除外したが，ここで，事例の数が少ないルールの確率値も，事例の数が多いルールと同じような確率分布を持っていると仮定すれば，ルールの確率の事前分布として，図 1 に示したような分布を用いることができる．また，証拠のタイプごとの性格を反映した分布にするために，本論文では，事前分布として，同一疑似単語中の同じタイプ（adjacent, window, pair）の証拠のルールの確率値の分布を用いることにする．

しかし，ベイズ推定において，直接，図 1 のような任意の分布を事前分布として用いると，事後分布が解析的に求まらなくなってしまう．そこで本論文では，実際に得られたルールの確率値の分布をベータ分布で近似することにする．ベータ分布は，ベルヌーイ試行において，自然共役事前分布と呼ばれる分布であり，事後確率の算

表 4: 事前分布の利用による確率値の推定精度向上の効果

品詞	疑似単語	adjacent	window	pair	all
名詞	政策／テレビ	86.2% (84.0%)	92.2% (88.4%)	82.0% (78.6%)	92.6% (90.1%)
	大統領／首相	82.2% (82.1%)	85.8% (84.6%)	83.2% (82.3%)	89.4% (88.2%)
	仕事／言葉／資金／文化	63.1% (65.5%)	66.6% (73.7%)	60.0% (64.5%)	75.1% (81.4%)
動詞	持つ／含む	81.8% (80.8%)	83.3% (81.4%)	87.3% (85.7%)	90.3% (89.4%)
	考える／見る／目指す	70.3% (64.5%)	63.2% (68.6%)	70.8% (69.4%)	73.6% (75.7%)
	入る／示す／開く／進める	61.7% (64.3%)	60.0% (68.6%)	76.6% (75.0%)	79.3% (82.1%)
形容詞	近い／難しい	89.6% (87.9%)	76.3% (77.7%)	85.7% (79.3%)	90.5% (90.7%)
	新しい／高い／強い	61.9% (62.8%)	67.8% (69.6%)	63.1% (63.8%)	72.6% (75.2%)
	若い／厳しい／大きい／よい	68.9% (67.2%)	66.7% (67.0%)	68.5% (65.5%)	77.4% (78.8%)
平均		74.0% (73.2%)	73.5% (75.5%)	75.2% (73.8%)	82.3% (83.5%)

(括弧内は判定に用いたルールの確率値の平均値)

出が解析的に簡単に求まることが知られている [9]. また，ベータ分布は，2つのパラメータを設定することで，非常に柔軟に様々な分布の形を表現できる．近似の方法としては，ベータ分布と，実際の確率値の分布の，平均と分散が等しくなるように2つのパラメータを設定した．図1の曲線は，図中の確率値の分布を近似したベータ分布である．

以下に事前分布をベータ分布で与えた場合の，事後分布の導出の過程を示す．まず，ベータ分布は次の式で与えられる．

$$P(\theta) = \frac{1}{B(a,b)}\theta^{(a-1)}(1-\theta)^{(b-1)} \tag{16}$$

この事前分布を式 (10) に代入することにより，事後分布は次のようになる．

$$P(\theta|A) = \frac{1}{B(a+k,b+n-k)}\theta^{(a+k-1)}(1-\theta)^{(b+n-k-1)}. \tag{17}$$

ただし，$B(a+k,b+n-k)$ はベータ間数である．

$$B(a+k,b+n-k) = \int_0^1 \theta^{(a+k-1)}(1-\theta)^{(b+n-k-1)} \tag{18}$$

事後分布の期待値は次のように得られる．

$$E[\theta] = \frac{a+k}{a+b+n}. \tag{19}$$

4.1 実験

表4に，事前分布としてベータ分布を用いた場合の結果を示す．事前分布を一様分布とした場合と比較すればわかるように，期待正解率と実際の正解率の差が小さくなっている．また，ルールの確率値の推定が正確になったことにより決定リストの性能が向上して判定の正解率も向上していることがわかる．

4.2 従来手法との正解率の比較

従来手法と正解率の比較を行なった．比較対象とした従来手法は，次の2つである．

- 間引き

  出現回数の少ない証拠のルールは使用しないようにする手法．出現回数が多い証拠であれば，最尤推定で求めた確率値を使用してもそれほど大きな誤差はないと期待される．しかし，この手法では，出現頻度の少ない証拠を捨ててしまうために，重要な証拠を見逃してしまう可能性がある．この実験では，出現回数が5回未満の事例のルールを削除した．ルールの確率値は式 (3) により算出した．

- 対数尤度比

  式 (2) を用いる手法．文献 [5][2] などで用いられている．この場合，式 (2) の分母が0になってしまう可能性があるため，頻度の比の式の分母と分子に小さな値 α を足す．このようにすることで，分母が0になってしまう問題を防げる．また，同じ確率であれば，頻度の高い証拠のルールが優先されるようになる．

表5に結果を示す．頻度の少ない事例のルールを間引く手法が最も正解率が低い，これはそのような間引きによって重要なルールが捨てられてしまうからだと考えられる．事前分布が一様分布によるベイズ推定と，対数尤度比を用いる手法は，ほぼ同じ正解率である．事前分布をベータ分布で近似した提案手法が，最も正解率が高くなっている．

表 5: 従来手法との正解率の比較

品詞	疑似単語	間引き	対数尤度比	ベイズ（一様分布）	ベイズ（ベータ分布）
名詞	政策／テレビ	88.3%	93.2%	92.0%	92.6%
	大統領／首相	83.4%	89.6%	88.6%	89.4%
	仕事／言葉／資金／文化	65.1%	72.1%	71.6%	75.1%
動詞	持つ／含む	84.5%	86.4%	86.6%	90.3%
	考える／見る／目指す	66.5%	65.4%	70.5%	73.6%
	入る／示す／開く／進める	72.2%	74.0%	76.3%	79.3%
形容詞	近い／難しい	84.5%	86.3%	88.6%	90.5%
	新しい／高い／強い	63.4%	68.7%	68.7%	72.6%
	若い／厳しい／大きい／よい	66.3%	71.8%	72.0%	77.4%
平均		74.9%	78.6%	79.4%	82.3%

5 関連研究

本論文では，ベイズ推定の枠組で証拠の種類ごとに異なった確率値を算出することで，決定リストの性能を向上させることができた．このように，証拠のタイプごとに，信頼度の値を変えることで決定リストの性能向上を図った研究として，文献 [2] がある．

文献 [2] では，決定リストによる同音異義語判別において，複合語からの証拠に重みを付けることで，分類精度の向上を図っている．そこでは，決定リストの信頼度として，式 (2) を用い，複合語からの証拠には，信頼度に重み付けのための係数を掛けることで，複合語からの証拠を用いたルールを優先させている．

本論文では，そのような，証拠の種類ごとに対する異なった重みづけをベイズ推定の枠組における事前分布を使用して行なったと考えることができる．本論文の手法には，次の 2 つの利点がある．一つは，どの種類の証拠に重み付けをするのかを，言語学的な直観に頼ることなく，実際のルールの確率値の分布によって決定できることである．もう一つは，どの程度重みを付けるのかに関しても，実際の確率値の分布を事前分布として利用することで，ベイズ推定の枠組から自動的に決定できるということである．

6 おわりに

本論文では，統計的クラス分類器である決定リストに対して，2 つの改善方法を示した．

- ベイズ推定によるルール確率値の推定

  決定リストを作成するにあたって最も重要なことは，ルールの信頼度をどのようにして計算するかということである．本論文では，ベイズ推定の手法を用いることにより，理論的な裏付けのあるスムージングによる推定が可能なことを示した．

- 証拠の種類ごとに事前分布を設定することによる精度向上

  証拠の種類ごとに，その性格を反映した事前分布にすることによって，より正確な確率値の推定ができることを示した．また，その結果，決定リストにおいて，より信頼性の高いルールが上位に位置するようになり，決定リストの性能が向上することを示した．

参考文献

[1] Hang Li and Kenji Yamanishi. Text classification using esc-based stochastic deicision lists. *Proc. of ACM-CIKM*, pp. 122–130, 1999.

[2] 新納浩幸. 複合語からの証拠に重みをつけた決定リストによる同音異義語判別. 情報処理学会論文誌, Vol. 39, No. 12, pp. 3200–3206, 1998.

[3] 新納浩幸. 日本語形態素解析のクラス分類問題への変換とその解法. 情報処理学会研究報告（2000-NL-135）, Vol. 2000, No. 11, pp. 149–156, 2000.

[4] 宇津呂武仁ほか. コーパスからの日本語従属節係り受け選好情報の抽出およびその評価. 自然言語処理, Vol. 6, No. 7, pp. 29–60, 1999.

[5] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, 1994.

[6] David Yarowsky. Hierarchical decision lists for word sense disambiguation. *Computers and Humanities*, pp. 34(2):179–186, 2000.

[7] 鈴木雪夫, 国友直人. ベイズ統計学とその応用. 東京大学出版会, 1989.

[8] 京都大学大学院情報学研究科. 日本語形態素解析システム JUMAN version 3.6, 1998.

[9] 繁桝算男. ベイズ統計入門. 東京大学出版会, 1985.