条件付き事前分布による文脈・状況に依存する先験知識の統合モデル

Integration model of context-specific and situated prior knowledge by conditional probability distributions

本村 陽一 †

Yoichi Motomura †

† 電子技術総合研究所

† Electrotechnical Lab, MITI.

Abstract: In the real world, information is sometimes observed incompletely. Then, an intelligent system working in the real environment has to combine observed information and prior knowledge of the environment. In order to realize this integration mechanism, a model based on Bayesian decision theory is proposed. In this framework, prior knowledge is represented as a conditional prior probability distribution, and integrated by Bayes' theorem. For optimum Bayesian decision theory based on a posterior probability, we have to approximate proper prior probabilities, which sometimes depends on particular situations. These situated prior probabilities can be modeled by Bayesian networks. Especially, Bayesian network on neural networks (BNNN) has some advantages for learning situated prior in the real world environment. As the demonstration of the framework, learning situated prior probabilities from an English dictionary is shown. It can apply to handwriting character recognition using Bayesian classifier with a neural network. The result improves recognition performance better than an ordinal (maximum likelihood) classifier.

1 はじめに

実環境における情報処理では限られた資源 (計算量・情報量) のもとでも頑健に動作することが求められる．例えば一度に得られるパターンが十分な情報を持っていないような場合であっても，我々人間はその状況や文脈に応じた適切な解を導くことができる場合がある．これはその環境固有の先見知識を用いることで観測データだけからは得られない情報を補っているためと考えられる．例えば人間はうす暗がりの中で十分な視覚情報が得られない時でも自分の部屋といった特定の環境であれば事前に獲得しているおおまかな位置関係の記憶をもとに，明示的な視覚に頼らなくても探しものを識別することができる．この様な先見知識と観測情報とを統合する枠組と，こうした先見知識を環境から学習によって獲得するメカニズムが実環境における頑健な知的システムのためには重要である．本研究ではこのような枠組の一つの例としてベイズの事後確率に基づく統合モデルを提案する．

連絡先: 〒305 つくば市梅園 1-1-4 電子技術総合研究所 情報科学部，本村 陽一, E-mail: motomura@etl.go.jp, Tel: (0298)54-5836, Fax: (0298)54-5301, URL: http://www.etl.go.jp/ motomura/

古くからベイズ的意志決定理論にしたがって人間の意志決定過程を説明する試みがあるが，実際的な場面における人間の直観的な解との乖離が心理学実験などを通じてしばしば指摘されてきた．しかし，最近これはベイズの定理の中で用いられる事前確率が固定的なものではなく状況や文脈に応じてダイナミックに変化するものと考え，これを忠実にモデル化することによりこの不整合が解決できるという主張がある [1]．またパターン認識においてベイズ事後確率に基づくベイズ識別器は理論上の最適解を与えるが，事後確率の計算中に必要となる事前確率を特定することが容易でないことからこれを一様分布とすることが多く行なわれている．この場合，事前確率は定数であるから単純に尤度の最大化をしていることになり，真の事後確率を最大にするとは限らない．そこで，識別時の状況やその文脈におけるパターンの出やすさといった先見知識をモデル化して事前確率を近似的に計算できれば，これを使った事後確率最大化によって従来の単純な尤度最大化に比べて性能が向上すると期待できる．この場合には先見知識と観測情報の統合はベイズの定理によって自然に行なわれる．

このような先見知識をベイズの事前確率として表現し実

際の状況や文脈に応じてどのようにモデル化するかが重要なポイントとなる． そこでこの事前確率を状況や文脈によって動的に変化する一種の条件付き確率であると考えこれを環境からの学習によって求めることを考える． 本研究ではこれまで筆者らが提案している神経回路上のベイジアンネット (BNNN: Bayesian network on Neural Networks) をこの問題に応用して，状況に依存した事前確率の学習を行なう．

以上の枠組の有効性を示すための簡単な例題としてアルファベット間の隣接関係についての先見知識を英単語辞書にある 4 万 3 千語のデータにより学習し，これと文字パターンの識別器として別に学習した 4 層の神経回路網と統合したモデルによって手書き文字認識を行なった結果を示す．

2 ベイズ的意志決定理論

ベイズ的な意志決定とは選択 $C = \{c_1, c_2, \cdots, \}$ についての重みづけを確率 $P(C)(\sum_i P(c_i) = 1)$ により行ない，情報 x を得た後での事後確率，$P(C|x)$ のもとでの期待効用，

$$U = \sum_i P(c_i|x)U(c_i) \tag{1}$$

を最大にするように決定を行なう手法とみなせる．

2.1 ベイズ識別

ベイズ的意志決定の簡単な例としてベイズ識別器を考える．すなわち，効用関数の値 $U(c_i)$ はどのクラス c_i についてもすべて等しいから，この場合はパターン x を観測した時の事後確率 $P(c_i|x)$ が最大となるクラス c_i にパターンを分類する．

ところでこの場合事後確率を直接計算することができないので，ベイズの定理を用いて，パターン x 尤度 $P(x|C)$ とクラスの事前確率 $P(C)$ を使って

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{\sum_i P(x|c_i)P(c_i)} \tag{2}$$

のように計算する． 尤度は観測入力 x から計算できるが，事前確率については一般的に偏りを持たないように無情報事前分布 (一様分布など) をとることが多く，この場合には尤度だけから決定を行なう最尤法と等価となる．

このベイズ識別器は理論上の誤り確率を最小にする決定であり，クラス間分散がクラス内分散よりも比較的大きい場合には図 1 のようになる．誤り率は二つの分布が重なりあう部分の面積に比例していて，この場合には比較的小さくなっている．しかし，クラス間分散がそれほど大きくない図 2 のような場合，分布の重なりが比較的大きく，誤り率もそれに比例して高くなる．

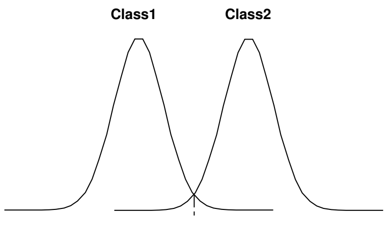

図 1: クラス間分散が大きい場合

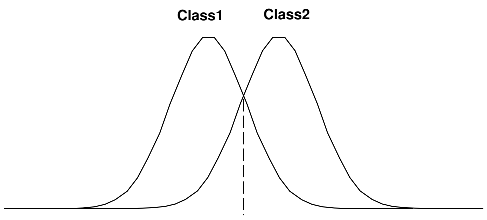

図 2: クラス間分散が小さい場合

これはそもそもデータ自体が似通っているため，尤度のみから識別することが無理な場合である． ところがこれは事前確率が一様と考えたことによる問題でもあり，もしも状況・文脈に依存して各クラスの発生頻度 (事前確率) に顕著な差があれば，より精密な事前確率を用いることによって誤り率を減少させることが期待できる．例えばクラス 1 とクラス 2 の発生頻度に差があり，それを事前確率として用いれば図 2 と同じクラス間分散であっても図 3，図 4 のように誤り確率を小さくすることができる．

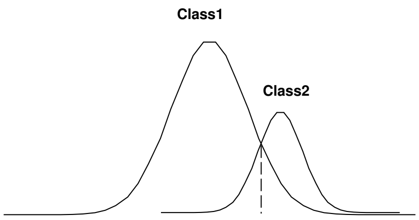

図 3: $P(Class1) > P(Class2)$ の場合

2.2 先見知識と観測情報との統合

実際に識別を行なう場合のパターンの発生頻度が事前分布 $P(C)$ と完全に合致している時にベイズ識別は理論上最適な結果を与えるはずである．そこで環境に存在している明示的には与えられない先見知識をモデル化し，学習により事前確率を近似する必要がある．実際の複雑な環境にお

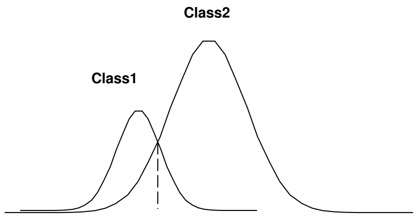

図 4: $P(Class1) < P(Class2)$ の場合

いてはパターンの発生過程に様々な必然があって，それにより事前確率は様々な状況や文脈に応じてある程度偏りがあることが普通である．つまり識別時の様々な状況や文脈を適切に切り分けて学習することでより精密な近似を行なうことができる．そこで事前分布は常に安定して存在しているものでではなく，状況や文脈に依存して決まる条件付き確率であるとして考える．例えばあるパタン x をクラス C のどれかに分類する時，実際にはクラス間の発生頻度が観測時の状況やパターンが存在している場所や前後の文脈に依存しているような場合，事前分布 $(P(C))$ は状況 (S) に依存したもの $(P(C|S))$ になっている．

この様な状況は有名な Selfridge のあいまいな文字列 [2](図 5) の場合にも見てとれる． 2 文字目はそれ自体だけではもはや'A' なのか'H' は判別がつかないが，1 文字目と 3 文字目の認識結果と我々が目にすることの多い単語による文字の並びのもっともらしさの統合の結果として上段は'H'，下段は'A' と結論づけられる．

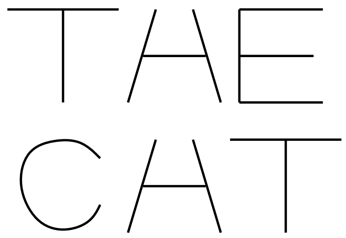

図 5: Ambiguous characters

この場合には隣接した 1 文字目と 3 文字目のパターンを状況 (あるいは文脈)S として，2 文字目のパターンの識別は S に依存した先見知識と統合されることで'A' か'H' かが決定されている．またこの先見知識は文字パターンとは明らかに別のレベルで抽象化されており，文字パターンが小文字であったり筆記体であったとしても同じ様に作用すると考えられる．そこで先見知識のモデル化はパターンの処理とは独立に対等に行なう．通常のパターン識別器は尤度 $P(x|c_i)$ を表すように構成されるが，これとは別に先見知識を条件付き確率 $P(c_i|S)$ として実際の各状況での発生頻度から学習する．

この両者を統合した式 (3) を最大とするクラスを識別結果とすることで我々が自然に行なっている認識過程が実現できる．

$$P(c_i|x,S) = \frac{P(x|c_i)P(c_i|S)}{\sum_i P(x|c_i)P(C_i|S)} \tag{3}$$

ここで分子第一項の $P(x|c_i)$ が観測情報に関する処理で，第二項 $P(c_i|S)$ が状況・文脈に依存する先見知識である．これらはベイズの定理の中の積演算として自然に統合されている．

3 ベイジアンネットによる先見知識のモデル化

前章で述べた状況に依存した先見知識を表す条件付き事前確率をどのように表現し，どのように学習するかがこの枠組をモデル化する上でのポイントとなる．本章以降では複雑な依存関係を持つ条件付き確率をグラフ構造で記述する確率モデル，ベイジアンネットによって環境内の状況・文脈に依存した先見知識を学習する方法を検討する．

3.1 グラフ構造を持つ確率モデル

各事象をノードで表し，その間の依存関係を矢印で結んでいくとある領域についての関連知識を非循環性の有向グラフ（Directed Acyclic Graph:DAG）として表すことができる．（$X \rightarrow Y$ において，X を親ノード，Y を子ノードと呼ぶ）この DAG と条件付き確率により構成される確率モデルが次に紹介するベイジアンネットである．確率変数を X，Y のように書きそれぞれの実現値を x, y とする．この X と Y の間に何らかの依存関係がある場合，例えば「もし $X = x$ ならば $Y = y$」というように記述することができる．しかし実世界の様々な事象を考えると依存関係はとても複雑に込み入っているため複雑な問題に対して「もし $X_1 = x_1, \cdots, X_i = x_i, \cdots$ ならば $Y = y$」のような決定論的な記述を完全に行なおうとすると記述量が膨大になりやがて破綻をきたす．そこで主たる依存関係だけに注目し，さらに不確実な事象をも扱うために「もし $X = x$ の時，$Y = y$ となる確率は $P(y|x)$」という確率化を導入する．この確率は $X = x$ となる状況のもとで，繰り返し観測できる事象 $Y = y$ の発生頻度，あるいは Y が y であると思われる主観的な確信度 (Degree of belief) として解釈される．

X と Y の全状態の起こり得る状態の発生頻度は条件付き確率分布 $P(Y|X)$ によって記述することができる． 領域内の全ての組合せ的な関係を記述するのでなく，明示的

な依存関係のない事象間は独立であると仮定し，主要な依存関係だけを抽出したグラフ構造を与えることにより計算量，記述量の爆発が避けられる．これはある種の近似的なモデルであり，例えば $A \cdots, Z$ の全状態の中で主たる依存関係が X, Y の間にだけ存在するような場合には全ての事象についての同時確率分布が次式右辺のように簡単化される．

$$\begin{aligned}&P(A, \cdots, X, Y, Z)\\&= P(A) \cdots P(X|Y)P(Y)P(Z)\end{aligned}$$

この様に ベイジアンネット による確率分布の表現は

- 直観的に捉えられた因果関係の間の構造（事象間の条件つき独立性）を容易に表現することができる．
- 確率的知識データの部分的な表現に適しており，他の領域，文脈へ転用する際にもメンテナンスしやすい．
- ルールベースで記述された古典的なエキスパートシステムに比べて直観的でわかりやすく，実環境での対話性に優れる．また局所的な計算のため並列分散化しやすい．逆に確率的なニューラルネットと比べた場合，学習によって獲得した知識は明示的に表現されるので取扱い易い．

といった特長がある．

3.2 神経回路上のベイジアンネット

前記の条件付き確率 $P(C|S)$ を表すためにベイジアンネットを用いると，複雑な依存関係にある確率変数の間の関係を有向グラフで与えることで様々な状況に応じた条件付き確率をモデル化することができる．特に条件付き確率を神経回路網で学習する BNNN を用いると連続値，多次元ベクトルを含むような多様な確率変数の取扱いが容易になるメリットがある

隠れ層を持つ多層の階層型ニューラルネットは任意の非線形写像を近似することが知られ，またそれに対する学習法である BP 法と合わせて様々な領域で応用されている．これをを各確率変数の条件付き確率の実現に用いることを考える．

筆者らは確率変数の次元に応じた入出力素子数を用意した一層の隠れ層を持つ次のような 3 層ニューラルネット $f(x)$ を用いて条件付き確率を表現するモデル，BNNN: Bayesian network on neural networks (神経回路の上に実現されたベイジアンネット) を提案している [9, 10]．これにより確率変数が離散多値，連続値の場合についても統一的に取扱え，学習も容易になるメリットがある．

Y が離散多値の確率変数 $Y = (y_1, y_2, \cdots, y_k)$ の場合，k 個のニューロン (クラスター) により Y の確率ベクトル $P(Y = y_1), P(Y = y_2), \cdots, P(Y = y_k)$ を表す．各確率変数に対応したニューロンクラスターを入力層，出力層として 1 つの隠れ層を持つ 3 層ニューラルネットを用いて $X = x$ のもとでの条件付き確率 $P(Y|X = x)$ を次のように表現する．

$$f_k(x) = g\left(\sum_j v_{jk} g\left(\sum_i w_{ij} x_i + b_j\right) + b_k\right)$$

$$g(x) = \frac{1}{1 + \exp(-x)}$$

$$P(y_k|x) = f_k(x) / \sum_k f_k(x)$$

Y が連続値の確率変数の場合，これの確率分布を μ, σ でパラメタライズしたガウス分布と考えて次のように表す．

$$f_\mu(x) = \sum_j v_j g\left(\sum_i w_{ij} x_i + b_j\right) + b_\mu \text{, } (f_\sigma\text{も同様})$$

$$P(y|x) = \frac{1}{\sqrt{2\pi} f_\sigma(x)} \exp\left(-\frac{(y - f_\mu(x))^2}{2 f_\sigma(x)^2}\right)$$

統計的に獲得した x,y のデーターセットから定まる条件付き確率を教師信号として与えることで，x を入力した時に $P(Y|x)$ が得られるようにニューラルネット，$f(x)$ を BP 法により学習する．

このモデルは 3 層の神経回路を多段に接続した階層型神経回路網の上に仮想的にベイジアンネットが表現されたものになっている．例えば確率変数 A, B, C, D の同時確率が

$$P(A, B, C, D) = P(D|C)P(C|A, B)P(B)P(A)$$

のような関係で表せる場合，図?? の左側のようなグラフでモデル化し，各変数の条件付き確率を学習した同図右側のニューラルネットにより実現できる．

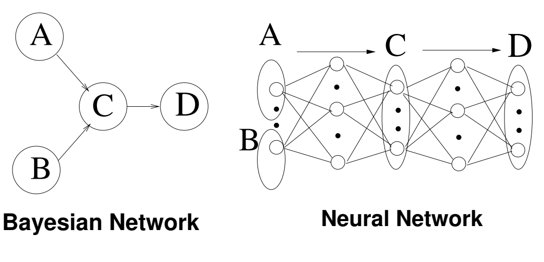

図 6: Bayesian network on neural networks

状況に依存する事前確率は状況 S をいくつかの確率変数の組で記述した条件付き確率をベイジアンネットで記述できれば，十分な実データを用いて学習できる．

4 応用

4.1 手書き文字認識のための統合システム

ここでは文脈・状況に依存する事前確率の学習と有効性を示すための比較的簡単な例題として手書き文字認識を考える．認識モジュールとして 20x20 のドットマトリクス上にマウスで書いた手書き英字 10 文字 {A,B,C,D,E,H,I,J,K,T} を 2x2 の受容野を持つ入力素子 (受容野中の濃度により $0 \cdots 4$ の値をとる) が 100 個，中間素子が 10 個の隠れ層 2 層, 出力素子が 10 個の計 4 層のニューラルネット, $g(x)$ を用いて図 7 の上二段の 10 パターンを平均二乗誤差が 0.01 以下になるまで学習して識別器を構成した．このニューラルネットの出力を文字パタン x を与えた時の文字 i に対応する近似尤度 $f_i(x)$ と書く．

一方，状況に依存する事前確率として，隣接する英字 C_{n-1}, C_{n+1} が与えられた時の間に挟まれた英字 C_n に関する条件付き確率 (10 次元確率ベクトル) $P(C_n|C_{n-1}, C_{n+1})$ を考え，4 万 3 千語からなる英単語辞書のデータを教師信号として BNNN(この場合，54-10-10 のニューラルネット) を学習した． この $S = (C_{n-1}, C_{n+1})$ を与えた時の $P(C_n|S)$ の確率ベクトルのうち文字 i に対応するものを $g_i(S)$ と書く．

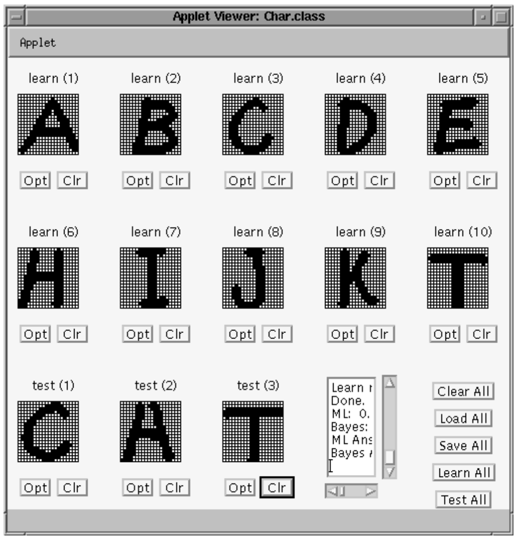

図 7: 上二段: 学習パターン, 下段: テストパターン

学習後に判読の困難な認識対象 (図 7 の下一段) を与え，次の I.(最尤解),II.(式 (3) に基づくベイズ解) の二者の比較を行なう．

I.$P(x|c_i) = f_i(x) / \sum_i f_i(x)$

II.$P(c_i|x, S) = P(x|c_i) \cdot g_i(S) / \sum_i (P(x|c_i) \cdot g_i(S))$

C_1 ='C', C_3 ='T' と認識した時の 2 文字目 C_2 について

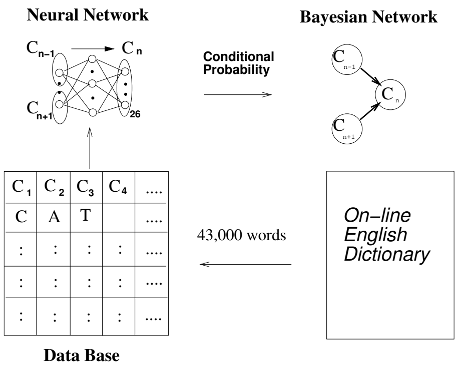

図 8: 学習の概要

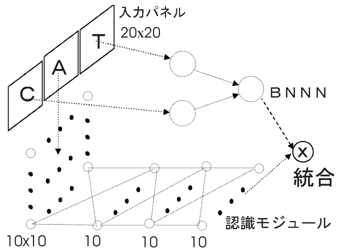

図 9: 実験システム概要

ての I., II. の推定結果を下表 (単位:%) に示す．

	A	B	C	D	E	H	I	J	K	T
I	15.6	3.3	17.9	9.1	7.5	22.1	8.6	3.8	9.8	2.1
II	75.2	0.0	0.4	0.0	3.3	4.7	12.7	0.0	3.8	0.0

これにより I. では誤認識されていた 2 文字目の判読困難な文字が，II. では 75.2% の高い確信度で 'A' と認識できている． 一方，C_1 ='T', C_3 ='E' の場合には C_2 ='H' となる推定結果は I. では 22.1% だったものが, II. では 88.5% であった．以上の実験から文脈・状況に依存する先見知識を統合したモデルによりより自然な認識結果が得られることがわかる．

5 議論

5.1 識別過程における先見知識と観測との相互作用

選択的注意機構のモデル [6] では認識モジュールと高次情報の制御回路が相互により密接に関わることで識別過程のダイナミクスが実現されている． 本論文，前章では 1 文字目と 3 文字目の識別が既に確定している場合を考えた

が, 識別過程が動的に進む場合は 1 文字目, 3 文字目の識別結果を確定せずベイズ事後確率による確信度を得ることに対応している. この時 1 文字目, 3 文字目の識別についてはその時点での二文字目の識別結果が影響すると考えればよい. 識別過程の t ステップ目の結果 (確信度) を C^t で表すと, 各ステップ毎に各文字の識別結果 C_i^t が交互に更新されていくことになる. したがって図 10 の様な時間発展を含むベイジアンネット (Temporal Dynamic Network) を用いて前章と同様の統合モデルを構成することができる.

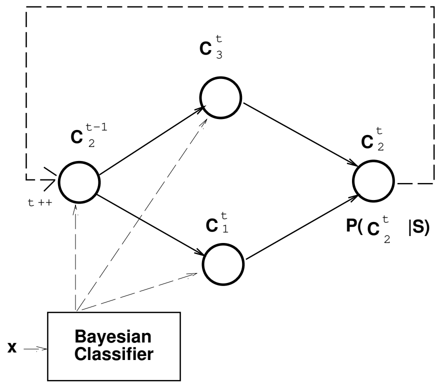

図 10: 認識過程における時間発展を含むベイジアンネット

5.2 関連研究

前後の依存関係を考慮して学習する認識モデルとして HMM(Hidden Markov Model) があり, 手書き認識へも適用されている [5] が, HMM では識別する単語毎にモデル化するのに対し, 本手法では文字間の文脈依存性と文字別器の学習を独立に行ない最終的に統合する点が大きく異なっている. 実際, 我々の先見知識は文字パターンとは明らかに別のレベルで抽象化されており, 文字パターンが小文字であったり筆記体であったとしても同じ様に作用すると考えられることから先見知識のモデル化をパターンの処理とは独立に対等に行なう本手法の方が自然な構成になっていると思われる. また実用化されている文字認識の中には単語辞書を用いるものも既に多くあるが, 本研究では辞書からは確率的な文字の隣接関係のみを先見知識として抽出しているため利用している情報, システムの性質が本質的に異なる.

5.3 今後の課題

記号とパターンの統合 [7, 8] という観点では, BNNN は任意の確率変数の相互関係についてはより抽象度の高いグラフ構造で表現できる特長がある. それを利用して今後は我々が身近に接しているより広範な問題について, 様々な文脈依存関係をグラフ構造で記述し認知的にもより人間と同様の先見知識を学習により獲得していくことが重要な課題である. 特に今回示した枠組はより広範なベイズ決定の応用問題にも適用できるため, 今後はカメラ画像からの物体認識についての実験評価なども行なう予定である.

6 謝辞

有益な御議論をいただいた電総研麻生英樹主任研究官 (学習統合基礎ラボリーダー) に感謝いたします. ホームページのアプレットを参考にさせていただいた慶応大学成田氏に深謝いたします.

参 考 文 献

[1] 繁桝: “意志決定の認知統計学”, 朝倉書店, (1995).

[2] Rumelhart,D.E., Hinton,G.E., and Williams,R.J.: “Learning internal represenations by error propagation,” in *Parallel Distributed Processing* Volume 1, J.L.McCleland, D.E.Rumelhard, and The PDP Ressearch group, MIT Press., (1986).

[3] 大津: “パターン認識と学習の基礎理論”, ニューロコンピューティングとその周辺, 脳とコンピュータ, 2, 培風館, (1995).

[4] H.Asoh: “Socially embedded learning of the office-conver robot Jijo-2”, Proc. of the 15th IJCAI, pp.880-885, (1997).

[5] Se-Chang Oh, Jin-Young Ha and Jin-H.Kim: “Context Dependent Search in Interconnected Hidden Markov Model for Unconstrained Hand Writing Recognition”, *Pattern Recognition*, Vol.28, No.11, pp.1693–1704, (1995).

[6] 庄野, 福島: “選択的注意機構を用いた英文筆記体文字列認識”, 神経回路学会第 6 回全国大会論文集, p.272–273, (1995).

[7] 石川, 松尾: “視覚認知におけるコンテクスト依存性の統合的モデル化の試み”, 記号とパターンの統合処理ワークショップ論文集, pp.9–14, (1994).

[8] 西, 大関, 桜井, 大森: “記号とパターンの統合による画像理解のための一方式”, 計測自動制御学会論文誌, vol.30, No.11, pp.1368–1377, (1994).

[9] Y.Motomura, I.Hara, H.Asoh, and T.Matsui: “Bayesian Network that Learns Conditional Probabilities by Neural Networks” , *Proc. of International Conference on Neural Information Processing '97* ,(1997).

[10] 本村 “多層パーセプトロンを用いたベイジアンネットによる確率推論”, 神経回路学会第 7 回全国大会論文集, p.243–244, (1996).