知能と情報（日本知能情報ファジィ学会誌）Vol.22, No.4, pp.485－492（2010）

解 説

ニューラルネットワーク最新事情(2)：ニューラルネットワークの数理†

櫻井 彰人*・篠沢 佳久*

1．前書き

戴いた題目はニューラルネットワークの数理であるが，「ニューラル」よりは多少広くとらえ，ネットワークを用いた計算モデルも紹介することにする．

ニューラルネットワークと呼ばれるものには非常に多くのバリエーションがある．本稿では，その中でもよく使われるフィードフォワード型を中心に記す．なお，一部ではあるが，フィードバック結線を付加したリカレント型についても記す．

機械学習の目的は，データを生成する情報源のある近似表現を求めることである．参照できるデータは有限量であり，またランダムにサンプルされたものであるため，真の表現を求めることは一般にはできない．真の表現に近い表現が得られる確率を高くするため，(残余誤差とのトレードオフを考えつつ)できるだけ簡単な表現を求めることが行われる．

情報源からのデータが，多様体や多様体で近似表現できる集合の要素のランダムな抽出であるなら，「簡単な表現」は「低次の多様体」と言うことができる．これまでの多くの研究によってニューラルネットワークが情報源の表現方法として有効であることが示されてきたが，それは，ニューラルネットワークはこの「低次の多様体」を仮説空間とする表現方法であり，そして，汎化能力が確保しやすい学習アルゴリズムが存在する表現方法であるためと推測することができる．なお，subspace learningについては，Neurocomputing, 73(10-12) が特集している．

ニューラルネットワークが広く研究されている理由には，機械学習器としての能力，すなわち，表現能力(できるだけ多様に表現できる)と汎化能力(限られたデータから情報源のできるだけ正確な表現が得られる)とを高めるという，相反する要求をよく実現していることが挙げられる．また，多くのバリエーションが考えられるという点も，重要である．後者においては他の学習器をはるかに凌駕している．

ニューラルネットワークは，その構造と学習の２つの側面から見ることができる．なおこの２点は，相互に関係しているため，どちらかに絞って話をすることは難しい．

ニューラルネットワークのパラメータ決定を機械学習の枠組みで考えたときには，構造は仮説空間に，学習アルゴリズムは探索バイアスに対応する．表現能力を確保するためには，構造を可変にする，パラメータ数を多くすることが考えられる．一方，汎化能力を確保するには探索アルゴリズムを適宜選んで，大きすぎる表現能力に振り回されなくする必要がある．そのために，実効的に，構造が単純なものやパラメータ数が少ないものを先に探索すればよい．しかし，一般にはそのような戦略をとることは容易ではない．

2．構造からみたニューラルネットワーク

構造を考える上での可変要素の主なものは，

・ 素子の入力関数・出力関数
・ 素子の結合状況を表すグラフ

である．

素子の入力関数は，線形関数とする．すなわち，入力値の重み付き線形和を関数値とする関数である．出力関数(以下では，正しくは活性化関数と出力関数の合成関数であるものを，出力関数と呼ぶ)は，この入力値の重み付き線形和に対して施され，素子の出力値を得る関数である(図１)．閾値関数(threshold function，step function)，標準シグモイド関数($\sigma(x)=1/(1+\exp(-x))$ または$\tanh(x)\equiv 2\sigma(x)-1$)を考える．

結合状況を表すグラフは，一般的にはdirected acyclic graphであるが，多くは，図２に示すように，結合関係に従いノードが層状に分類できるグラフを用いている．入力層には実質的な働きを持った素子はなく，単に端子である．これに対し中間層と出力層は実質的な役目をもつ．中間層は，入力層や出力層と異なり，外部から直接アクセスできないため，隠れ層とも呼ばれる．層を飛び越える結合を許すこともある．層

† Neural Networks and their Mathematical Aspects
Akito SAKURAI and Yoshihisa SHINOZAWA

* 慶應義塾大学理工学部
Faculty of Science and Technology, Keio University

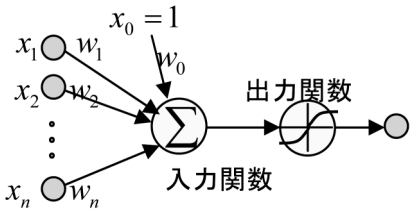

図1　素子

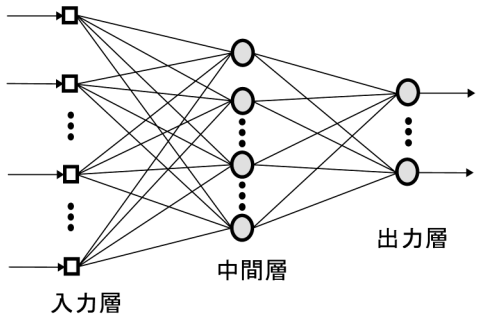

図2　層状構造

を越えた結合の有無の表現力に対する影響は，理論的解析からはわかっていない．実際に学習させたときの効果は，データに依存する．

典型的な構造として，ノード1個，中間層1層，中間層2層以上のものを考えることができるが，本稿では，中間層1層の場合についてのみ記す．なお，中間層数nのものを，n+2層ネットワークということが多い．これは，計算量理論での数え方では，深さn+1である．

2.1　分類問題における汎化能力

構造の違いが生み出す表現力の差を，分類問題については，VC次元を用いて評価することができる（[Sontag 1998]）．VC次元は仮説空間の表現能力を測る道具であり，次のように定義される（[Vapnik and Chevonenkis 1971]，[Blumer et al. 1989]）．

定義：仮説空間HのVC次元 $VC\dim(H)$は，Hによってshatterされうる点の個数の最大値である．ある点集合Sが仮説空間Hによりshatterされるというのは，Sのどのような2分割に対しても，それと整合的な仮説がHの中にあることである．

VC次元を用いて，汎化誤差を評価することができる．例えば，次の定理がある．

定理：仮説空間H中のデータに整合的な仮説hに対し，確率$1-\eta$で次式が成立する．

$$R(h) \le R_{emp}(h) + \sqrt{\left(d\left(\ln 2N/d + 1\right) - \ln \eta/4\right)/N}$$

ただし，$d = VC\dim(H)$とし，$R(h)$，$R_{emp}(h)$はそれぞれ，汎化誤差，学習誤差であり，それぞれ，

$$R(h) = \int (1/2)\left|h(x) - y\right| dP(x,y)\ ,$$

$$R_{emp}(h) = (1/2N)\sum_{i=1}^{N}\left|h(x_i) - y_i\right| \text{ である．}$$

VC次元を用いた汎化誤差の評価は非常に大雑把なため実用には適さないが，傾向を示していると考える．なお，この大雑把さは，学習データの分布に関する事前知識を仮定しないためである．事前知識を仮定する研究もある．

以下では，入力空間はn次元ユークリッド空間$\Re^n$，データ数はN，中間素子数はhで表す．

中間層一層の場合を記す（[Sontag 1998]，[Sakurai 1993，1995]，[Macintyre and Sontag 1993]）．閾値素子を用いた場合のVC次元については，$\Theta(nh\log h)$が得られている．標準シグモイド素子を用いた場合については，そのVC次元は，$\Omega(nh\log h)$かつ$O((nh)^2h^2)$である．中間層2層以上については，閾値素子を用いた場合はほぼ，$\Theta(((1/2)h^2+nh)\log h)$，標準シグモイド素子の場合は，$\Omega(((1/2)h^2+nh)\log\ h)$かつ$O((nh)^2h^2)$である．

2.2　近似問題における汎化能力

近似問題に関しては，universal approximation theoremが得られている（[Funahashi 1989]，[Cybenko 1989]）．

定理：$\phi(\cdot)$を定数でない，非有界な単調増加関数とする．$f(\cdot)$を，$\Re^n$のあるコンパクト集合K上の実数値連続関数とする．任意の$\varepsilon>0$に対して，$|f(x)-f_N(x)|<\varepsilon$となるように，整数$N$，実定数$\theta_i$，$w_{ij}$と$f_N(x)=\sum_{i=1}^{N} c_i\phi\left(\sum_{j=1}^{n} w_{ij}x_j - \theta_i\right)$を定めることができる．但し，$x=(x_1, \ldots, x_n)$である．

これは重要な結果であるが，この定理及び証明からは，実際にどの程度の近似精度が得られるかは分からない．実際，$f(\cdot)$が連続とはいえ，変化が大きい点が多く存在すれば必要とするNはいくらでも大きくなり

うる．Barronは，関数$f(\cdot)$にある制約を加えることにより，汎化誤差が$O(1/n)$とできることを示した．

定理[Barron 1993]：関数$f(x)$は$f(x)=\int_{\Re^n} e^{i\omega\cdot x}\tilde{f}(\omega)d\omega$とFourier変換でき，しかも，$C_f=\int_{\Re^n}|\omega|\left|\tilde{f}(\omega)\right|d\omega$が有限値であるとする．このとき，$B_r=\{x\,|\,|x|\le r\}$上の任意の確率測度$\mu$に対して$\int_{B_r}\left(f(x)-f_N(x)\right)^2\mu(dx)\le(2rC_f)^2/N$とできる．

3．学習

ニューラルネットワークの場合には，分類問題・近似問題どちらを対象とするにせよ，仮説空間が広く，(誤差の定め方にも依存するが)類似仮説が多く存在する．しかし，それら仮説の汎化能力にはばらつきがある．そのため実際の学習アルゴリズムを構成するには，

・誤差関数
・探索方法

の2つを適切に決定する必要がある．とはいえ，探索順序の制御は困難であり，通常は適用する最適化アルゴリズム依存である．適切な汎化能力を持たせるためには，むしろ，誤差関数に正則化項を付加する(加算する)ことが選択される．

3.1 入出力値と誤差関数

入力値に制限はなく，従って，$\Re^n$内の任意の点を入力とする．出力値は，出力層素子の出力関数に依存する．例えば，線形素子であれば，通常は$\Re^m$の任意の値をとることができ，すなわち，$\Re^m$の任意の値を教師信号とすることができる．しかし，標準シグモイド関数の場合は，各素子の出力値が有界となり，異なる解釈となる．

$\sigma(x)$が出力関数の場合を例にする．離散値を教師信号とする場合を考える．出力値が開区間$(0,1)$のため，教師信号を0と1とする方法もあるが，通常は，$\varepsilon>0$を適宜選択し，εと$1-\varepsilon$を用いる．またこの場合，確率値と解釈することもある．

誤差関数としては，2乗誤差関数，log-loss関数が用いられることが多い．2乗誤差関数は出力値が正規分布に従うときにのみ正当化されるが，通常はそれを検証せずに用いている．出力値を確率として解釈するときには，log-loss関数を用いるのが普通である．

誤差関数を設定し，そのグローバルな最小値が求めうるのであれば，学習アルゴリズムの良し悪しはその収束速度(と必要な記憶容量)によって判断される．しかし，局所最小値に囚われるおそれがあったり，繰返し計算途中での打ち切りを考慮する場合には，学習アルゴリズムに内在する仮説空間の探索順序をも考慮する必要がある．

3.2 Perceptron learning algorithm

線形閾値素子1個の場合の学習アルゴリズムとしては，Rosenblattのperceptron learning algorithmが有名であり，また，多くの場合有効である．入力結合荷重ベクトルwの更新式は$w\leftarrow w+\eta(t-o)x$となる．ただし，tは目標出力値，oは実際の出力値(いずれも離散値)，ηは学習係数である．学習係数は正定数であればなんでもよい．

学習データが線形分離可能であれば必ず収束することが知られている．しかし，線形分離可能でない場合は収束せず，ある有界集合内を循環する．

Rosenblattの更新式は実用にはあまり用いられなかったが，Freund and Schapireがlarge margin classifierを得るのに，これを変形したvoted perceptron learning algorithmを提案した以降，研究が進められている．CCG (combinatory categorial grammar) のパーザーであるC&C parserの学習に用いた例が報告されている([Clark and Curran 2007])．

閾値素子一個では表現能力が限られる．すなわち，線形分離可能な関数しか実現できない．しかし，これを多層化すれば，例えば中間層1層設ければ(本来のmultilayer perceptron)，素子数を必要に応じて増やすことにより，任意の論理関数が実現できる．しかし，中間素子が分担して解消すべき誤差量を，中間素子ごとに適切に定めることが難しく(こうした問題は一般には，credit assignment問題と呼ばれる)，適切な計算量の学習アルゴリズムを構築することはできなかった．確率的アルゴリズムは可能であるがその性質は分かっていない[Sakurai 2001]．

3.3 連続な出力関数

素子の関数や誤差関数を微分可能な関数とすれば，最急降下法や他の最適化アルゴリズムが適用可能である．なお，最急降下法($w\leftarrow w-\eta\partial E/\partial w$)のように勾配に基づく方法を，一般に，勾配法と呼ぶ．

なかでも最急降下法であるBP(back propagationまたはerror back propagation)がよく知られている．中間素子が分担して解消すべき誤差量が，出力値を計算する時とは逆順に，すなわち出力素子から入力素子へ

と，計算することにより求められる．これが名称の由来である．

BPは簡便であり，実際，容易な課題，小規模なネットワーク（中間層数１，中間素子数で10程度）では容易に収束することが知られている．しかし，課題が困難であったりすると（例えば，排他的論理和であっても入力数が数個でも），収束が遅くまたは収束しなくなる．

収束が遅いのは，誤差関数の変化が緩やかな場所が広いため（plateauと呼ばれる）と考えられる．そこで，例えば，慣性項を用いる方法 $w^{(k+1)}=w^{(k)}-\eta\partial E/\partial w+\alpha(w^{(k)}-w^{(k-1)})$ が提案され用いられた．しかし，（局所）最適点付近での振舞いが不安定になるという副作用があり，必ずしも薦められる方法ではない．なお，Quickprop, Rprop等多数のよりよい改善方法が提案されている．

こうしたplateauの原因の一つに，中間層素子の入れ替え対称性がある．勾配法を用いると，この対象性に起因する鞍点によりplateauが発生すると推定される．

自然勾配法（natural gradient algorithm）は，パラメータ w のFisher情報行列の逆行列 G を用いて，$w\leftarrow w-\eta G^{-1}\partial E/\partial w$ とする方法である（[Amari 1998]）．パラメータ空間をリーマン空間とみた場合，通常の勾配方向は最急方向ではなく，自然勾配方向が最急方向であり，それはFisher情報行列により定まる．自然勾配法は，上記鞍点付近ではplateauに会うことなく学習が進むと考えられる．なお，鞍点は特異点でもあるため学習が不安定になるとも考えられるが，実際には，安定していることが示されている．

最近では，Levenberg-Marquardt法（[Masters 1995]）やL-BFGS（[Apostolopoulou et al. 2009]）が高速に安定して収束するとして，用いられている．実際的な問題を解くときには，これらを用いるのが薦められる．

4．いろいろなネットワーク

4.1 Recurrent Neural Network

時系列の学習を行うために，中間層の出力値や出力層の出力値を入力層にフィードバックするネットワークである．フィードバックするにあたって，一時点遅らせる．これにより，階層型ネットワークと同じ学習アルゴリズムで，離散時間の時系列を学習させることができる．

最も単純なものは，中間層出力をそのまま（但し一時点遅れ）入力層に追加するものであり，SRN（simple recurrent network）と呼ばれ，文法の学習を試みたことで知られる（[Elman 1990]）．

4.2 Echo state network

ニューラルネットワークの学習能力の高さから，例えば，chaotic sequenceの学習が可能かどうかに興味がある．残念ながら通常のRNNではうまくいかない．

一つの方法としてEcho state networkが提案されている（[Jaeger 2004]）．これは，相互結合を含む，大規模かつランダムな入力層・中間層間の固定結合（適度にスパースかつ学習しない）をもつ．中間層・出力層間の結合荷重は最小二乗法で定める．これにより高速かつ正確にカオス時系列が学習できる．もっとも，学習するというより，予測精度の高いサブネットワークの選択を行っているといえる．

なお，liquid state machineも類似のアイデアに基づくものである［Maass et al. 2002］．

4.3 中間層における学習

機械学習は低次元表現を求めることと解釈できると最初に述べた．ニューラルネットワークで低次元表現を明示的に求めるのが，bottleneck neural network（砂時計型ニューラルネットワーク）である．

図３のような中間層一層の標準シグモイド関数素子のネットワークを考える．この場合の仮説空間は，関数 $f_1:\Re^n\mapsto\Re^h$ と $f_2:\Re^h\mapsto\Re^n$ の合成関数 $f_2\circ f_1$ の集合である．標準シグモイド関数を用いた場合，$f_2\circ f_1\equiv I$（恒等写像）とはできないため，情報源が h 次元多様体で表されても，当該ニューラルネットワークは近似表現を実現するのみである．

しかし，入力を，低次元の中間層を通して出力層に再現するためには，中間層で効率の良いデータ表現を行う必要がある．すなわち，情報圧縮を行う必要がある．その結果，中間層には圧縮した表現が得られる．

こうした現象は，通常のfeedforwardやrecurrent

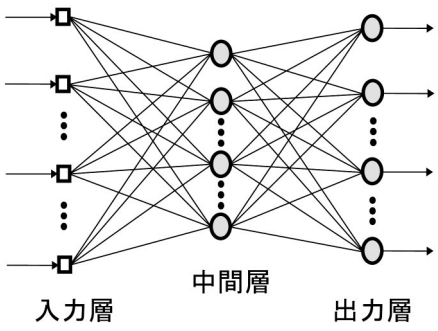

図３　砂時計ネットワーク

ネットワークでも発生する．入力データごとの中間層出力値のクラスタリングを行うと，入力値→出力目標値の変換方法が近いと思われる入力値に対応する中間層出力値が同一・近隣のクラスタに属するようになる．

リカレント型で時系列の予測学習をさせると，n-gramを学習したようなパターンが出現することが知られている（[Pearlmutter 1989]）．Elmanは，SRNに文法学習を行わせ，使用方法の近い単語が，中間素子出力値のなす空間上で，近い位置に配置されるように学習されることを示した（[Elman 1990]）．

4.4 主成分分析（PCA）

主成分分析（PCA）がニューラルネットワークで行える．前節で述べた情報圧縮が，主成分で行われる場合があることを利用する．

通常は線形素子を用いる．最も単純にはHebb学習則を用いる．Hebb則は，神経素子が行っている学習方法として最も初期に推測・提案されたものであり，$w \leftarrow w+\eta x(w \cdot x)$とするものである．

ただし，Hebb則を直截に適用したのでは一般には発散するため，wのベクトル長を正規化する．[Oja 1982]は，これに対し，$w \leftarrow w+\eta x(w \cdot x)-\alpha(w \cdot x)^2 w$という改善を提案した．右辺最後の項は$w$の発散を防ぐ忘却項である．いずれの場合も，$w$に第一主成分が得られる．

複数個の主成分を求めるネットワークとしては，Ojaの方法の多要素版である subspace modelやadaptive principal components extraction（APEX）が知られている．前者は，個々の主成分を求めるものではなく，それが張る空間を張るベクトルを求める．後者は，側抑制を用いて個々の主成分を求める．

なお，bottleneckネットワークでも，主成分が張る空間を張るベクトルが求まる．入力値と出力値の教師値を同一として，学習させる．線形素子と二乗誤差を用いれば，最初のh個の主成分が入力層と中間層間の結合荷重に表れる．なお，中間層と出力層間の結合荷重は，入力層・中間層間の結合荷重行列の転置行列となる．

なお，独立成分分析を行うネットワークを構成することができる．[Jutten and Herault 1991]，[Amari et al. 1995]を例としてあげておく．

4.5 強化学習

強化学習は，あるグラフ上をエージェントが永遠に動きまわるとして，エージェントが獲得する利得（reinforcement）を最大にする方策（次に移るノードを示す指示書．確率的でもよい）を学習する方法を与える．イメージとしては，テキスト・アドベンチャー（interactive fictionともいう）を考えればよい（実際に強化学習を用いた研究がある）．

状態を表す変数をノードとし，状態遷移が可能な方向を有向エッジとする有向グラフを考える．エッジや状態に利得が割り当てられている（利得の付与は確率的でよい）．ゴール状態に到達する方策を学習させたい場合，しばしば，ゴール状態のみに利得をおき，ゴールに達成したエージェントを初期状態に戻すという枠組みを用いる．永遠に動き回るとすると獲得利得の累積は発散するので，獲得利得を経過時間に従って割り引いた割引獲得利得を考えたり，初期状態からゴールまでを１つのエポックとしてその間の獲得利得を考えたり，割り引かずに獲得利得を経過時間で平均した平均利得を考えたりする．

グラフと利得関数が既知であれば，動的計画法により最適方策を求めることができる．強化学習の眼目は，グラフも利得も未知であるときに，探索しながら，最適方策を得る方法を与える点にある．学習アルゴリズムは，状態または状態・遷移対に（最適方策のもと得られる）期待獲得利得を与える関数を，学習する．

探索（試行錯誤）時に得た，状態または状態・遷移対と利得との関係を記録するだけでは，学習効率が悪すぎる．ある時に得た利得を，それに至る探索系列に従って，過去に巡った状態または状態・遷移対に，分配する必要がある．これもcredit assignment問題である．

この問題に対し，シグモイド素子をもつ多層ニューラルネットワークの場合には，最終誤差を，合成関数の微分に従い，前の層の素子の誤差に換算していた．強化学習においても類似のアイデアに基づく．強化学習の学習方法としては，TD学習，その基礎となるvalue/policy iteration，その発展形等が多く提案され

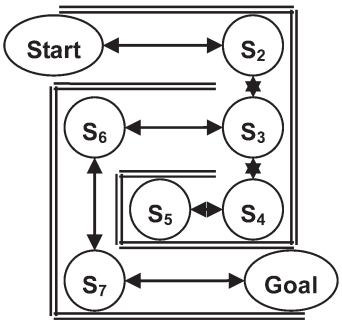

図 4 迷路探索時の状態

ている．

強化学習でも，非常に多くのパラメータを学習する必要がある．強化学習アルゴリズムには，獲得報酬の割引や利得の分配を通じた，内在的なスムージングが組み込まれている．しかし，ニューラルネットワークを用いれば，より広い範囲でスムージングができ，そして次元の呪いが回避できる可能性がある．Neuro-Dynamic Programmingはその一例である（[Bertsekas and Tsitsiklis 1996]，[Tesauro 1992]）．ただし，学習アルゴリズムには工夫が必要である．

4.6 Bayesian network

Bayesian networkは，確率変数間の「因果関係」を有向グラフで表現したものである．「因果関係」と書いたが，勿論，これは間違いである．確率変数を用いては，外在的関係として定義しない限り，因果関係・原因結果関係は表現できない．相関関係は表現できるので，Bayesian networkでは，（データまたは事前知識により）相関関係の存在が否定できない確率変数間をエッジで結んだ（確率変数がノードである）無向グラフを，まず考える．その上で，何らかの事前知識により変数間には順序（因果関係に整合的な順序）が定義されているとして，その順序に整合的にエッジの方向を決める．

Bayesian networkは，従って，確率変数の結合確率を，ある変数間の相関関係のなさを用いて，簡潔に表現しようとしたものといえる．確率変数x_i間の順序を$x_i \rightarrow x_{i+1}$とし，確率変数x_iへ入るエッジの他端を$\pi(x_i)$で表せば，$P=(x_1, \ldots, x_n)=\prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$

$=\prod_{i=1}^{n} P(x_i \mid \pi(x_i))$と表現できる．ただし，$\pi(x_i)$が空の場合，$P(x_i \mid \pi(x_i))=P(x_i)$とする．多くの場合確率変数は離散値をとるものとして，各確率変数（ノード）には，条件付確率表$P(x_i \mid \pi(x_i))$を持たせる．

Bayesian networkでは，変数値がすべて与えられた場合の結合確率を，親のないノードから始めて，有向エッジに従い，結合確率を順番に求めることにより高速に求めることができる（ニューラルネットワークのforward propagationに相当する）．実は，確率変数に値が定まっていないものがあるときには，その分布を他の変数の条件付確率表として，高速に求めるアルゴリズムが構築されている（backpropagationに相当するといいたいが，少々異なる）．

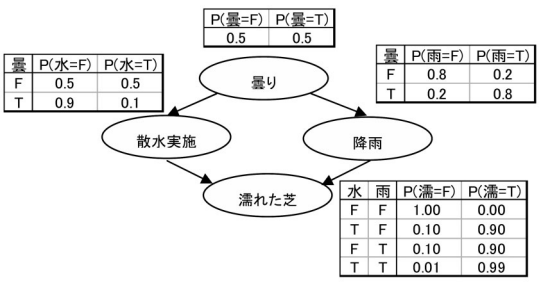

図5 Bayesian networkの例

Bayesian networkの学習は，構造であるグラフの学習と条件付確率表の学習とからなる．条件付確率表を用いるのは，各確率変数は多項分布に従うと仮定するからであり，これは基本的には，各変数値の組の発生頻度から求める（最尤推定）．スムージングには，共役分布であるDirichlet分布を事前分布として用いる．構造学習には，何らかの情報量規準を用い，ヒューリスティックを用いた網羅的探索を行う．

ところで，条件付確率表の推定においては，分布形を予め定め，当該変数に関する分布のみを考えている．全変数間の影響が反映された，分布形を想定しない学習方法があってしかるべきである．実際，ニューラルネットワークによる表現と学習を用いた学習方法が提案されている（[Motomura 2001]）．

5．終わりに

本稿では，ニューラルネットワークに関し，普段はあまり注目されないが重要な内容である，構造と汎化能力について記した後，ネットワークに基づく計算機構について，ほんの一部を記した．最近では，ニューラルネットワークとは離れるかとも思うが，グラフィカルモデルやlatent variableを用いた様々なモデル（特にDirichlet過程に基づくもの）が提案され，その有効性が示されている．本稿を，そういったものに対するアプローチのきっかけとして戴ければ幸いである．

参 考 文 献

[Amari 1998] Amari, S. Natural gradient works efficiently in learning. Neural Computation, 10:251－276, 1998.

[Amari et al. 1995] Amari, S., A. Cichocki, and H. H. Yang. Recurrent Neural Networks for Blind Separation of Sources. Proceedings of International Symposium on Nonlinear Theory and Applications, NOLTA '95 Las Vegas, NV, Dec. 10－14, 1995.

[Apostolopoulou et al. 2009] M.S. Apostolopoulou, D.G. Sotiropoulos, I.E. Livieris and P. Pintelas, A Memoryless BFGS Neural Network Training Algorithm, Proceedings of 6th IEEE International Conference on Industrial Informatics（INDIN 2009）, pp.216－221, 2009.

[Barron 1993] Barron, A. R.: Universal approximation

bounds for superpositions of a sigmoidal function. IEEE Transsactions on Information Theory, 36:930－945, 1993.

［Bertsekas and Tsitsiklis 1996］ Bertsekas, D. P. and J. N. Tsitsiklis, Neuro－Dynamic Programming. Athena Scientific, Bellmont, MA, 1996.

［Blumer et al. 1989］ Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik－Chervonenkis dimension. Journal of the ACM, 36（4）: 929－865, 1989.

［Clark and Curran 2007］ Clark, S. and J. R. Curran. Perceptron Training for a Wide－Coverage Lexicalized－Grammar Parser, Proceedings of the Workshop on Deep Linguistic Processing, pp.9－16, 2007.

［Cybenko 1989］ Cybenko, G. Approximations by superpositions of sigmoidal functions. Mathematics of Control, Signals, and Systems, 2 : 303－314, 1989

［Elman 1990］ Elman, J.L. Finding Structure in Time. Cognitive Science 14:179－211, 1990.

［Freund and Schapire 1999］ Freund, Y. and R. E. Schapire. Large margin classification using the perceptron algorithm. Machine Learning 37（3）: 277－296, 1999.

［Funahashi 1989］ Funahashi, K. On the approximate realization of continuous mappings by neural networks, Neural Networks, 2 : 183－192, 1989.

［Jaeger 2004］ Jaeger, H.. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science, 304 : 78－80, 2004.

［Jutten and Herault 1991］ Jutten, C., J. Herault : Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture, Signal Processing, 24 : 1－20, 1991.

［Kurkova 1992］ Kurkova V. Kolmogorov's Theorem and multilayer neural networks, Neural Networks, 5 : 501－506, 1992.

［Maass et al. 2002］ Maass, W., T. Natschlaeger, and H. Markram. Real-time computing without stable states : a new framework for neural computation based on perturbations, Neural Computation, 14（11）: 2531－2560, 2002.

［Macintyre and Sontag 1993］ Macintyre, A. and E.D. Sontag. Finiteness results for sigmoidal neural networks. In STOC '93 : Proceedings of the twenty－fifth annual ACM symposium on Theory of computing, pages 325－334, 1993.

［Masters 1995］ Masters, T. Advanced Algorithms for Neural Networks : A C++ Sourcebook, New York, NY : John Wiley & Sons, 1995.

［Minsky and Papert 1969］ Minsky M. L. and S. A. Papert. Perceptrons. MIT Press, 1969.

［Motomura 2001］ Motomura, Y. BAYONET : Bayesian Network on Neural Network, Foundations of Real-World Intelligence, pp.28－37, CSLI publications, Stanford, Calfornia, 2001.

［Oja 1982］ Oja, E. A simplified neuron model as a principal component analyzer., Journal of Mathematical Biology, 15 : 267－273, 1982.

［Oja 1989］ Oja, E. Neural networks, principle components, and subspaces. International Journal of Neural Systems, 1 : 61－68, 1989.

［Pearl 1998］ Pearl, J., Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco, CA, 1988.

［Pearlmutter 1989］ Pearlmutter, B. A. Learning state space trajectories in recurrent neural networks. Neural Computation, 1（2）: 263－269, 1989.

［Rosenblatt 1958］ Rosenblatt, Frank（1958）, The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain, Psychological Review, 65（6）: 386－408, 1958.

［Sakurai 1993］ Sakurai, A. Tighter Bounds of the VC－Dimension of Three-layer Networks, Proceedings of the World Congress on Neural Networks 1993, vol.3, pp.540－543.

［Sakurai 1995］ Sakurai, A. Polynomial bounds for the VC－dimension of sigmoidal, radial basis function, and sigma－pi networks, Proceedings of World Congress on Neural Networks 1995, vol.I, 58－63.

［Sakurai 2001］ Sakurai, A. A Fast and Convergent Stochastic MLP Learning Algorithm, International Journal of Neural Systems, 11（6）: 573－584, 2001.

［Sontag 1998］ Sontag, E. D. VC dimension of neural networks. In C.M. Bishop, editor, Neural Networks and Machine Learning, pages 69－95. Springer, Berlin, 1998.

［Tesauro 1992］ Tesauro, G. Practical Issues in Temporal Difference Learning, Machine Learning, 8 : 257－277, 1992.

［Vapnik and Chevonenkis 1971］ Vapnik, V. and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 16（2）: 264－280, 1971

（2010年７月７日　受付）

［問い合わせ先］
〒223-8522　横浜市港北区日吉３-14-１
慶應義塾大学理工学部
櫻井　彰人
TEL：045－563－1141（代）
FAX：045－566－1617

著 者 紹 介

櫻井 彰人［非会員］

1975年東京大学工学部計数工学科卒業，77年同大学院情報工学研究科修了．77年日立製作所計測器事業部那珂工場．78年 Dept. Computer Science, Univ. Illinois（MsCS取得(1979)）．89年同基礎研究所，96年同中央研究所，98年北陸先端科学技術大学院大学教授を経て，2002年慶應義塾大学理工学部教授となる．博士（工学）（東京大学1993）．人工神経回路網，機械学習を専門とする．

篠沢 佳久［非会員］

1994年慶應義塾大学理工学部管理工学科卒業，96年同大学院理工学研究科修士課程修了，99年同研究科博士課程修了．博士（工学）．現在慶應義塾大学理工学部専任講師．パターン認識，人工神経回路網に関する研究に従事．