(19)日本国特許庁(JP)　(12)特　許　公　報(B1)　(11)特許番号
特許第6054005号
(P6054005)
(45)発行日　平成28年12月27日(2016.12.27)　(24)登録日　平成28年12月9日(2016.12.9)

(51)Int.Cl.　F I
G06N 3/04 (2006.01)　G06N 3/04

請求項の数 18　(全 27 頁)

(21)出願番号　特願2016-548332(P2016-548332)
(86)(22)出願日　平成27年8月31日(2015.8.31)
(86)国際出願番号　PCT/JP2015/074720
審査請求日　平成28年7月25日(2016.7.25)
(31)優先権主張番号　特願2015-113440(P2015-113440)
(32)優先日　平成27年6月3日(2015.6.3)
(33)優先権主張国　日本国(JP)

早期審査対象出願

(73)特許権者　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目７番３号
(74)代理人　100123434
弁理士　田澤　英昭
(74)代理人　100101133
弁理士　濱田　初音
(74)代理人　100199749
弁理士　中島　成
(74)代理人　100188880
弁理士　坂元　辰哉
(74)代理人　100197767
弁理士　辻岡　将昭
(74)代理人　100201743
弁理士　井上　和真

最終頁に続く

(54)【発明の名称】推論装置及び推論方法

(57)【特許請求の範囲】
【請求項１】
ニューラルネットワークを構成している入力層の各ノードにデータが与えられると、前記データから前記入力層の各ノードでの活性度を算出する入力層活性度算出部と、
前記ニューラルネットワークを構成している中間層のノードと前記入力層のノードとを接続している各エッジの重みを記憶している中間層記憶部と、
前記入力層活性度算出部により算出された入力層の各ノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記中間層の各ノードと接続関係がある前記入力層の各ノードでの活性度と各エッジの重みとを取得し、前記取得した入力層の各ノードでの活性度と各エッジの重みとを用いて、前記中間層の各ノードでの活性度を算出する中間層活性度算出部と、
前記中間層活性度算出部により算出された中間層の各ノードでの活性度を用いて、前記ニューラルネットワークを構成している出力層の各ノードでの活性度を算出する出力層活性度算出部と
を備えており、かつ、中間層の各ノード当りの前記入力層との平均接続本数が５０本以下であることを特徴とする推論装置。
【請求項２】
前記ニューラルネットワークを構成している中間層が複数存在しており、
前記中間層記憶部は、前記ニューラルネットワークを構成している中間層毎に、当該中間層のノードが前記入力層のノードと接続されていれば、当該中間層のノードと前記入力

層のノードとを接続している各エッジの重みを記憶し、当該中間層のノードが他の中間層のノードと接続されていれば、当該中間層のノードと他の中間層のノードとを接続している各エッジの重みを記憶しており、

前記中間層活性度算出部は、前記ニューラルネットワークを構成している中間層のノードと接続されているノードが前記入力層のノードであれば、前記入力層活性度算出部により算出された入力層の各ノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記ニューラルネットワークを構成している中間層の各ノードと接続関係がある前記入力層の各ノードでの活性度と各エッジの重みとを取得し、前記ニューラルネットワークを構成している中間層のノードと接続されているノードが他の中間層のノードであれば、他の中間層のノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記ニューラルネットワークを構成している中間層の各ノードと接続関係がある他の中間層のノードでの活性度と各エッジの重みとを取得し、前記取得した入力層又は他の中間層のノードでの活性度と各エッジの重みとを用いて、前記ニューラルネットワークを構成している中間層の各ノードでの活性度を算出し、かつ、全部または一部の中間層の各ノード当りの前記中間層との平均接続本数が５０本以下であることを特徴とする請求項１記載の推論装置。

【請求項３】

前記出力層のノードと前記中間層のノードとを接続している各エッジの重みを記憶している出力層記憶部を備え、

前記出力層活性度算出部は、前記中間層活性度算出部により算出された中間層の各ノードでの活性度及び前記出力層記憶部に記憶されている各エッジの重みの中から、前記出力層の各ノードと接続関係がある前記中間層の各ノードでの活性度と各エッジの重みとを取得し、前記取得した中間層の各ノードでの活性度と各エッジの重みとを用いて、前記出力層の各ノードでの活性度を算出することを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項４】

前記出力層のノードが前記入力層のノードと接続されていれば、前記出力層のノードと前記入力層のノードとを接続している各エッジの重みを記憶し、前記出力層のノードが前記中間層のノードと接続されていれば、前記出力層のノードと前記中間層のノードとを接続している各エッジの重みを記憶している出力層記憶部を備え、

前記出力層活性度算出部は、前記出力層のノードと接続されているノードが前記入力層のノードであれば、前記入力層活性度算出部により算出された入力層の各ノードでの活性度及び前記出力層記憶部に記憶されている各エッジの重みの中から、前記出力層の各ノードと接続関係がある前記入力層の各ノードでの活性度と各エッジの重みとを取得し、前記出力層のノードと接続されているノードが前記中間層のノードであれば、前記中間層活性度算出部により算出された中間層のノードでの活性度及び前記出力層記憶部に記憶されている各エッジの重みの中から、前記出力層の各ノードと接続関係がある中間層のノードでの活性度と各エッジの重みとを取得し、前記取得した入力層又は中間層のノードでの活性度と各エッジの重みとを用いて、前記出力層の各ノードでの活性度を算出することを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項５】

前記中間層記憶部は、前記各エッジの重みのほかに、前記中間層の各ノードに与えられているバイアス値を記憶しており、

前記中間層活性度算出部は、前記入力層の各ノードでの活性度と各エッジの重みと前記バイアス値とを用いて、前記中間層の各ノードでの活性度を算出することを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項６】

前記中間層活性度算出部は、前記中間層のノード毎に、当該ノードと接続関係がある前記入力層の各ノードでの活性度と、前記中間層の当該ノードと前記入力層の各ノードを接続している各エッジの重みとの積和演算を実施して、前記積和演算の演算結果と前記中間

層の当該ノードのバイアス値を加算し、その加算した結果を前記ニューラルネットワークの活性化関数の引数として用いることで、前記中間層の当該ノードの活性度として、前記活性化関数の関数値を算出することを特徴とする請求項５記載の推論装置。

【請求項７】

前記出力層記憶部は、前記各エッジの重みのほかに、前記出力層の各ノードに与えられているバイアス値を記憶しており、

前記出力層活性度算出部は、前記中間層の各ノードでの活性度と各エッジの重みと前記バイアス値とを用いて、前記出力層の各ノードでの活性度を算出することを特徴とする請求項３記載の推論装置。

【請求項８】

前記出力層活性度算出部は、前記出力層のノード毎に、当該ノードと接続関係がある前記中間層の各ノードでの活性度と、前記出力層の当該ノードと前記中間層の各ノードを接続している各エッジの重みとの積和演算を実施して、前記積和演算の演算結果と前記出力層の当該ノードのバイアス値を加算し、その加算した結果を前記ニューラルネットワークの活性化関数の引数として用いることで、前記出力層の当該ノードの活性度として、前記活性化関数の関数値を算出することを特徴とする請求項７記載の推論装置。

【請求項９】

前記ニューラルネットワークを構成している入力層及び中間層のノードを接続している複数のエッジ、あるいは、前記ニューラルネットワークを構成している中間層及び出力層のノードを接続している複数のエッジがループをなしており、前記ループが６本以上のエッジで形成されていることを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項１０】

前記ニューラルネットワークを構成している複数の中間層のノードを接続している複数のエッジがループをなしており、前記ループが６本以上のエッジで形成されていることを特徴とする請求項２記載の推論装置。

【請求項１１】

前記ニューラルネットワークを構成している中間層の各ノードは、前記入力層における全ノードのうち、ランダムに選択された一部のノードと接続されており、

前記ニューラルネットワークを構成している出力層の各ノードは、前記中間層における全ノードのうち、ランダムに選択された一部のノードと接続されていることを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項１２】

前記ニューラルネットワークを構成している中間層の各ノードは、前記入力層又は他の中間層における全ノードのうち、ランダムに選択された一部のノードと接続されていることを特徴とする請求項２記載の推論装置。

【請求項１３】

前記ニューラルネットワークを構成している中間層の各ノードは、前記入力層における全ノードのうち、隣接していない一部のノードと接続されており、

前記ニューラルネットワークを構成している出力層の各ノードは、前記中間層における全ノードのうち、隣接していない一部のノードと接続されていることを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項１４】

前記ニューラルネットワークを構成している複数の中間層の各ノードは、前記入力層又は他の中間層における全ノードのうち、隣接していない一部のノードと接続されていることを特徴とする請求項２記載の推論装置。

【請求項１５】

前記ニューラルネットワークを構成している中間層の各ノード当りの前記入力層のノードとの平均接続本数が前記入力層のノードの個数の１０分の１以下であることを特徴とする請求項１あるいは請求項２記載の推論装置。

【請求項１６】

前記ニューラルネットワークを構成している複数の中間層の各ノード当りの前記入力層又は他の中間層のノードとの平均接続本数が前記入力層又は他の中間層のノードの個数の１０分の１以下であることを特徴とする請求項２記載の推論装置。

【請求項１７】

中間層記憶部が、ニューラルネットワークを構成している中間層のノードと入力層のノードとを接続している各エッジの重みを記憶しており、

入力層活性度算出部が、前記ニューラルネットワークを構成している入力層の各ノードにデータが与えられると、前記データから前記入力層の各ノードでの活性度を算出し、

中間層活性度算出部が、前記入力層活性度算出部により算出された入力層の各ノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記中間層の各ノードと接続関係がある前記入力層の各ノードでの活性度と各エッジの重みとを取得し、前記取得した入力層の各ノードでの活性度と各エッジの重みとを用いて、前記中間層の各ノードでの活性度を算出し、

出力層活性度算出部が、前記中間層活性度算出部により算出された中間層の各ノードでの活性度を用いて、前記ニューラルネットワークを構成している出力層の各ノードでの活性度を算出し、かつ、中間層の各ノード当りの前記入力層との平均接続本数が５０本以下であることを特徴とする推論方法。

【請求項１８】

前記ニューラルネットワークを構成している中間層が複数存在しており、

前記中間層記憶部は、前記ニューラルネットワークを構成している中間層毎に、当該中間層のノードが前記入力層のノードと接続されていれば、当該中間層のノードと前記入力層のノードとを接続している各エッジの重みを記憶し、当該中間層のノードが他の中間層のノードと接続されていれば、当該中間層のノードと他の中間層のノードとを接続している各エッジの重みを記憶しており、

前記中間層活性度算出部は、前記ニューラルネットワークを構成している中間層のノードと接続されているノードが前記入力層のノードであれば、前記入力層活性度算出部により算出された入力層の各ノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記ニューラルネットワークを構成している中間層の各ノードと接続関係がある前記入力層の各ノードでの活性度と各エッジの重みとを取得し、前記ニューラルネットワークを構成している中間層のノードと接続されているノードが他の中間層のノードであれば、他の中間層のノードでの活性度及び前記中間層記憶部に記憶されている各エッジの重みの中から、前記ニューラルネットワークを構成している中間層の各ノードと接続関係がある他の中間層のノードでの活性度と各エッジの重みとを取得し、前記取得した入力層又は他の中間層のノードでの活性度と各エッジの重みとを用いて、前記ニューラルネットワークを構成している中間層の各ノードでの活性度を算出することを特徴とし、かつ、全部または一部の中間層の各ノード当りの前記中間層との平均接続本数が５０本以下であることを特徴とする請求項１７記載の推論方法。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は、ニューラルネットワークを用いた推論装置及び推論方法に関するものである。

【背景技術】

【０００２】

機械学習手法の１つとして、ニューラルネットワークは高い問題解決能力を有し、画像認識、音声認識、異常検知、将来予測などの多くの処理に用いられている。

ニューラルネットワークの構造の１つとして、階層型ニューラルネットワークがあり、学習手法として、主に教師あり学習と、教師なし学習の二種類がある。

教師あり学習は、複数の学習例の入力データと目標出力を与え、実際の出力と目標出力が一致するように、ニューラルネットワークの結合状態を調整する手法である。また、教

師なし学習は、目標出力を与えずに、学習例が有する本質的な特徴を抽出できるように、ニューラルネットワークの結合状態を調整する手法である。

【０００３】

例えば、教師あり学習法に属する誤差逆伝播法（バックプロパゲーションアルゴリズム）には、ニューラルネットワークの階層の数が多くなると、学習結果が収束しなくなる問題を発生することがある。

上記の問題を解決するために、例えば、自己符号化器（Ａｕｔｏｅｎｃｏｄｅｒ）や制約ボルツマンマシン（Ｒｅｓｔｒｉｃｔ　Ｂｏｌｔｚｍａｎｎ　Ｍａｃｈｉｎｅ）などの教師なし学習を用いて層毎の事前学習（Ｐｒｅ－ｔｒａｉｎｉｎｇ）を実施することで、ニューラルネットワークの結合状態の初期値を決定し、その後、誤差逆伝播法を用いて、ニューラルネットワークの結合状態を調整（Ｆｉｎｅ－ｔｕｎｉｎｇ）するようにしているものがある。

これにより、学習結果が収束しなくなる問題の発生を招くことなく、実際の出力と目標出力が一致するように、ニューラルネットワークの結合状態を調整することができる。

【０００４】

階層型ニューラルネットワークは、複数のノード（節点）及びノード間の結合するエッジ（枝）で構成されるグラフ構造で表すことができるが、例えば、４層のニューラルネットワークでは、複数のノードが入力層、第１中間層、第２中間層、出力層で階層化され、同一階層に属するノード間のエッジは存在せずに、隣接している層の間だけにエッジが存在する。中間層は隠れ層と呼ばれることがある。

各エッジには、繋いだ２つのノード間の結合度合を示すパラメータが存在し、そのパラメータは、エッジ重みと呼ばれている。

【０００５】

階層型ニューラルネットワークを用いて、学習または推論を行う際、その計算量及びメモリ量はエッジ数に比例する。一般的に、各層に属するノードは、隣接している層に属する全てのノードとエッジで接続されているため、計算量及びメモリ量がノード数と直接に関係する。

例えば、入力層のノード数がＮ、第１中間層のノード数が$Ｍ_{1}$、第２中間層のノード数が$Ｍ_{2}$、出力層のノード数が１である場合には、入力層と第１中間層の間のエッジ数がＮ×$Ｍ_{1}$、第１中間層と第２中間層の間のエッジ数が$Ｍ_{1}$×$Ｍ_{2}$、第２中間層と出力層の間のエッジ数が$Ｍ_{2}$となるため、学習または推論を行う際の計算量及びメモリ量が、（Ｎ×$Ｍ_{1}$＋$Ｍ_{1}$×$Ｍ_{2}$＋$Ｍ_{2}$）に比例する。

【０００６】

特に、中間層のノード数が入力層のノード数に比例する場合、ノード数がＮ個の入力層に対して、第１中間層のノード数が$Ｍ_{1}$＝ａ×Ｎ個、第２中間層のノード数が$Ｍ_{2}$＝ｂ×Ｎ個となる。この場合、ニューラルネットワークにおけるエッジの総数がＮ×ａ×Ｎ＋ａ×Ｎ×ｂ×Ｎ＋ｂ×Ｎ＝（ａ＋ａ×ｂ）×$Ｎ^{2}$＋ｂ×Ｎになり、学習または推論を行う際の計算量及びメモリ量が、（ａ＋ａ×ｂ）×$Ｎ^{2}$＋ｂ×Ｎに比例する。

【０００７】

階層型ニューラルネットワークは、以上のような構造を持つことが多く、計算量及びメモリ量が、入力データ数であるＮの２乗、即ち、入力層のノード数Ｎの２乗に比例して増加するため、入力データ数の増大と共に計算量及びメモリ量が飛躍的に増加し、計算機リソース不足、処理遅延、装置コスト増大などの問題が発生する。

以下の特許文献１には、複数の入力データの相関関係に基づいて、複数の入力データをグループ化することで、入力層と中間層の間のエッジ数や、中間層と出力層の間のエッジ数を削減している。

【先行技術文献】

【特許文献】

【０００８】

【特許文献１】特開２０１１－５４２００号公報（図１）

【発明の概要】
【発明が解決しようとする課題】
【０００９】
従来の推論装置は以上のように構成されているので、入力層と中間層の間のエッジ数や、中間層と出力層の間のエッジ数を削減することができる。しかし、同一グループに属する入力層と中間層の間では、入力層の各ノードが、中間層の全てのノードと接続されるため、エッジの削減数が限定的であり、依然として、推論を行う際の計算量及びメモリ量が大きくなってしまうという課題があった。
【００１０】
この発明は上記のような課題を解決するためになされたもので、推論を行う際の計算量及びメモリ量を削減することができる推論装置及び推論方法を得ることを目的とする。また、推論精度が高い推論装置及び推論方法を得ることを目的とする。
【課題を解決するための手段】
【００１１】
この発明に係る推論装置は、ニューラルネットワークを構成している入力層の各ノードにデータが与えられると、そのデータから入力層の各ノードでの活性度を算出する入力層活性度算出部と、ニューラルネットワークを構成している中間層のノードと入力層のノードとを接続している各エッジの重みを記憶している中間層記憶部と、入力層活性度算出部により算出された入力層の各ノードでの活性度及び中間層記憶部に記憶されている各エッジの重みの中から、中間層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとを取得し、その取得した入力層の各ノードでの活性度と各エッジの重みとを用いて、中間層の各ノードでの活性度を算出する中間層活性度算出部と、中間層活性度算出部により算出された中間層の各ノードでの活性度を用いて、ニューラルネットワークを構成している出力層の各ノードでの活性度を算出する出力層活性度算出部とを備えており、かつ、中間層の各ノード当りの前記入力層との平均接続本数が５０本以下であることを特徴とするようにしたものである。
【発明の効果】
【００１２】
この発明によれば、推論を行う際の計算量及びメモリ量を削減することができる効果がある。また、他の効果として、より高い推論精度を得ることができる。
【図面の簡単な説明】
【００１３】
【図１】この発明の実施の形態１による推論装置を示す構成図である。
【図２】この発明の実施の形態１による推論装置を示すハードウェア構成図である。
【図３】推論装置がコンピュータで構成される場合のハードウェア構成図である。
【図４】この発明の実施の形態１による推論装置の処理内容である推論方法を示すフローチャートである。
【図５】入力層活性度算出部１、第１中間層活性度算出部５、第２中間層活性度算出部６、第３中間層活性度算出部７及び出力層活性度算出部９の処理内容を示すフローチャートである。
【図６】この発明の実施の形態１による推論装置が適用する階層型ニューラルネットワークを示す説明図である。
【図７】インデックス、エッジ重み及びバイアス値の一例を示す説明図である。
【図８】複数のエッジによって形成されるループの例を示す説明図である。
【図９】入力層に与えられる画像データを１０種のクラスに識別するニューラルネットワークの一例を示す説明図である。
【図１０】入力層から第１中間層、第１中間層から第２中間層（文脈層）、第２中間層（文脈層）から第１中間層、第１中間層から出力層に枝接続を持っているリカレントニューラルネットワーク（ＲＮＮ）であるエルマンネット（Ｅｌｍａｎ　Ｎｅｔｗｏｒｋ）を示す説明図である。

【図１１】この発明の実施の形態７による推論装置を示す構成図である。
【図１２】中間層内でノード同士の枝接続や自己接続が存在するとともに、入力層から中間層を飛ばして出力層へ接続する枝が存在するニューラルネットワークであるＥｃｈｏ　Ｓｔａｔｅ　Ｎｅｔｗｏｒｋの例を示す説明図である。
【図１３】図１２のＥｃｈｏ　Ｓｔａｔｅ　Ｎｅｔｗｏｒｋを層単位で表している説明図である。
【図１４】この発明の実施の形態８による推論装置を示す構成図である。
【発明を実施するための形態】
【００１４】
以下、この発明をより詳細に説明するために、この発明を実施するための形態について、添付の図面にしたがって説明する。
【００１５】
実施の形態１．
図１はこの発明の実施の形態１による推論装置を示す構成図であり、図２はこの発明の実施の形態１による推論装置を示すハードウェア構成図である。
図１では、複数のノードが、入力層、第１中間層、第２中間層、第３中間層、出力層で階層化されている５層の階層型ニューラルネットワークを用いる推論装置の例を示している。また、図１の例では、入力層に与えられるデータが画像データである例を示している。
ここでは、５層の階層型ニューラルネットワークを用いる例を示しているが、５層の階層型ニューラルネットワークに限るものではなく、３層や４層、あるいは、６層以上の階層型ニューラルネットワークを用いるものであってもよい。
因みに、３層の階層型ニューラルネットワークを用いる場合、中間層は第１中間層だけになり、後述する第２中間層記憶部３、第３中間層記憶部４、第２中間層活性度算出部６及び第３中間層活性度算出部７が不要になる。
【００１６】
図１及び図２において、入力層活性度算出部１は例えばＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）を実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている入力層活性度算出回路１１で実現されるものであり、階層型ニューラルネットワークを構成している入力層の各ノードに画像データが与えられると、その画像データから入力層の各ノードでの活性度を算出する処理を実施する。
【００１７】
中間層記憶部である第１中間層記憶部２は例えばＲＡＭやハードディスクなどの記憶媒体からなる中間層記憶装置１２で実現されるものであり、第１中間層の各ノードと入力層の各ノードとの接続関係を示すインデックス（接続情報）と、第１中間層のノードと入力層のノードとを接続している各エッジの重みと、第１中間層の各ノードに与えられているバイアス値とを記憶している。
中間層記憶部である第２中間層記憶部３は例えばＲＡＭやハードディスクなどの記憶媒体からなる中間層記憶装置１２で実現されるものであり、第２中間層の各ノードと第１中間層の各ノードとの接続関係を示すインデックスと、第２中間層のノードと第１中間層のノードとを接続している各エッジの重みと、第２中間層の各ノードに与えられているバイアス値とを記憶している。
中間層記憶部である第３中間層記憶部４は例えばＲＡＭやハードディスクなどの記憶媒体からなる中間層記憶装置１２で実現されるものであり、第３中間層の各ノードと第２中間層の各ノードとの接続関係を示すインデックスと、第３中間層のノードと第２中間層のノードとを接続している各エッジの重みと、第３中間層の各ノードに与えられているバイアス値とを記憶している。
【００１８】
中間層活性度算出部である第１中間層活性度算出部５は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出回

路１３で実現されるものであり、第１中間層記憶部２に記憶されているインデックスを参照して、入力層活性度算出部１により算出された入力層の各ノードでの活性度と第１中間層記憶部２に記憶されている各エッジの重み及びバイアス値の中から、第１中間層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した入力層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第１中間層の各ノードでの活性度を算出する処理を実施する。

【００１９】

中間層活性度算出部である第２中間層活性度算出部６は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出回路１３で実現されるものであり、第２中間層記憶部３に記憶されているインデックスを参照して、第１中間層活性度算出部５により算出された第１中間層の各ノードでの活性度と第２中間層記憶部３に記憶されている各エッジの重み及びバイアス値の中から、第２中間層の各ノードと接続関係がある第１中間層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した第１中間層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第２中間層の各ノードでの活性度を算出する処理を実施する。

【００２０】

中間層活性度算出部である第３中間層活性度算出部７は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出回路１３で実現されるものであり、第３中間層記憶部４に記憶されているインデックスを参照して、第２中間層活性度算出部６により算出された第２中間層の各ノードでの活性度と第３中間層記憶部４に記憶されている各エッジの重み及びバイアス値の中から、第３中間層の各ノードと接続関係がある第２中間層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した第２中間層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第３中間層の各ノードでの活性度を算出する処理を実施する。

【００２１】

出力層記憶部８は例えばＲＡＭやハードディスクなどの記憶媒体からなる出力層記憶装置１４で実現されるものであり、出力層の各ノードと第３中間層の各ノードとの接続関係を示すインデックス（接続情報）と、出力層のノードと第３中間層のノードとを接続している各エッジの重みと、出力層の各ノードに与えられているバイアス値とを記憶している。

出力層活性度算出部９は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている出力層活性度算出回路１５で実現されるものであり、出力層記憶部８に記憶されているインデックスを参照して、第３中間層活性度算出部７により算出された第３中間層の各ノードでの活性度と出力層記憶部８に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある第３中間層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した第３中間層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、出力層の各ノードでの活性度を算出する処理を実施する。

【００２２】

図１では、推論装置の構成要素である入力層活性度算出部１、第１中間層記憶部２、第２中間層記憶部３、第３中間層記憶部４、第１中間層活性度算出部５、第２中間層活性度算出部６、第３中間層活性度算出部７、出力層記憶部８及び出力層活性度算出部９のそれぞれが専用のハードウェアで構成されているものを想定しているが、推論装置がコンピュータで構成されていてもよい。

図３は推論装置がコンピュータで構成される場合のハードウェア構成図である。

推論装置がコンピュータで構成される場合、第１中間層記憶部２、第２中間層記憶部３、第３中間層記憶部４及び出力層記憶部８をコンピュータのメモリ２１上に構成するとともに、入力層活性度算出部１、第１中間層活性度算出部５、第２中間層活性度算出部６、第３中間層活性度算出部７及び出力層活性度算出部９の処理内容を記述しているプログラムをコンピュータのメモリ２１に格納し、当該コンピュータのプロセッサ２２がメモリ２

１に格納されているプログラムを実行するようにすればよい。
図４はこの発明の実施の形態１による推論装置の処理内容である推論方法を示すフローチャートであり、図５は入力層活性度算出部１、第１中間層活性度算出部５、第２中間層活性度算出部６、第３中間層活性度算出部７及び出力層活性度算出部９の処理内容を示すフローチャートである。

【００２３】

図６はこの発明の実施の形態１による推論装置が適用する階層型ニューラルネットワークを示す説明図である。
図６の階層型ニューラルネットワークは、複数のノードが、入力層、第１中間層、第２中間層、第３中間層、出力層で階層化されている５層の階層型ニューラルネットワークである。
図７はインデックス、エッジ重み及びバイアス値の一例を示す説明図である。
図７では、ノードの接続情報であるインデックスが、例えば、第１中間層における“Ｎ”のノードは、入力層における“０”，“３”，“５”のノードと接続されている旨を示している。
また、図７では、例えば、第１中間層における“Ｎ”のノードと、入力層における“０”のノードとを接続しているエッジの重みが“０．２”、第１中間層における“Ｎ”のノードと、入力層における“３”のノードとを接続しているエッジの重みが“－０．５”、第１中間層における“Ｎ”のノードと、入力層における“５”のノードとを接続しているエッジの重みが“０．１”である旨を示している。
さらに、図７では、例えば、第１中間層における“Ｎ”のノードのバイアス値が“１．８”である旨を示している。

【００２４】

次に動作について説明する。
入力層活性度算出部１は、階層型ニューラルネットワークを構成している入力層の各ノードに画像データが与えられると、その画像データから入力層の各ノードでの活性度A_{IN}を算出する（図４のステップＳＴ１）。
入力層活性度算出部１に与えられる画像データが、例えば、０～２５５の画素値Ｐを有するピクセルからなる画像を示すデータであり、各ピクセルの画素値Ｐが入力層の各ノードに与えられる場合、入力層の各ノードでの活性度A_{IN}は、下記の式（１）のように算出することができる。

$$A_{IN}=\frac{P}{255} \qquad (1)$$

ここでは、画像データが入力される場合を想定し、各ピクセルの画素値Ｐを２５５で除算することで正規化して、浮動小数点値（０．０～１．０）を入力層の各ノードでの活性度A_{IN}とする例を示しているが、単なる正規化のみでなく、入力されるデータの種類に応じて、データ間引き、量子化、変換等の処理を実施するようにしてもよい。

【００２５】

第１中間層活性度算出部５は、入力層活性度算出部１が入力層の各ノードでの活性度A_{IN}を算出すると、第１中間層記憶部２に記憶されているインデックスを参照して、第１中間層のノード毎に、当該ノードに接続されている入力層の各ノードを確認して、その入力層の各ノードでの活性度A_{IN}を取得する。
例えば、第１中間層における“Ｎ”のノードの場合、第１中間層記憶部２に記憶されているインデックスが、入力層における“０”，“３”，“５”のノードと接続されている旨を示しているので、入力層活性度算出部１により算出された入力層の各ノードでの活性度A_{IN}のうち、入力層における“０”，“３”，“５”のノードでの活性度A_{IN-0}，A_{IN-3}，A_{IN-5}を取得する。

【００２６】

また、第１中間層活性度算出部５は、第１中間層記憶部２に記憶されているインデックスを参照して、第１中間層のノード毎に、当該ノードに接続されているエッジを確認して

、第１中間層記憶部２からそのエッジの重みｗを取得する。
例えば、第１中間層における“Ｎ”のノードの場合、第１中間層記憶部２に記憶されているインデックスが、入力層における“０”，“３”，“５”のノードと接続されている旨を示しているので、第１中間層における“Ｎ”のノードと、入力層における“０”のノードとを接続しているエッジの重みw_{N-0}として“０．２”を取得し、第１中間層における“Ｎ”のノードと、入力層における“３”のノードとを接続しているエッジの重みw_{N-3}として“－０．５”を取得する。また、第１中間層における“Ｎ”のノードと、入力層における“５”のノードとを接続しているエッジの重みw_{N-5}として“０．１”を取得する。

【００２７】

また、第１中間層活性度算出部５は、第１中間層のノード毎に、第１中間層記憶部２から当該ノードのバイアス値B_{1M}を取得する。
例えば、第１中間層における“Ｎ”のノードの場合、バイアス値B_{1M-N}として“１．８”を取得する。

【００２８】

第１中間層活性度算出部５は、第１中間層のノード毎に、入力層の各ノードでの活性度A_{IN}、エッジの重みｗ、バイアス値B_{1M}を取得すると、活性度A_{IN}、エッジの重みｗ、バイアス値B_{1M}を用いて、第１中間層のノード毎の活性度A_{1M}を算出する（ステップＳＴ２）。
以下、第１中間層における“Ｎ”のノードの活性度A_{1M-N}の算出例を具体的に説明する。

【００２９】

まず、第１中間層活性度算出部５は、第１中間層記憶部２に記憶されているインデックスを読み込んで（図５のステップＳＴ１１）、そのインデックスを参照することで、活性度の算出に用いるパラメータとして、入力層における“０”，“３”，“５”のノードでの活性度A_{IN-0}，A_{IN-3}，A_{IN-5}と、エッジの重みw_{N-0}，w_{N-3}，w_{N-5}と、第１中間層における“Ｎ”のノードのバイアス値B_{1M-N}とを取得する（ステップＳＴ１２）
次に、第１中間層活性度算出部５は、下記の式（２）に示すように、入力層における“０”，“３”，“５”のノードでの活性度A_{IN-0}，A_{IN-3}，A_{IN-5}と、エッジの重みw_{N-0}，w_{N-3}，w_{N-5}との積和演算を実施する（ステップＳＴ１３）。

$$MADD = A_{IN-0} \times w_{N-0} + A_{IN-3} \times w_{N-3} + A_{IN-5} \times w_{N-5} \quad (2)$$

次に、第１中間層活性度算出部５は、下記の式（３）に示すように、積和演算の演算結果ＭＡＤＤと第１中間層における“Ｎ”のノードのバイアス値B_{1M-N}とを加算する（ステップＳＴ１４）。

$$ADD = MADD + B_{1M-N} \quad (3)$$

【００３０】

第１中間層活性度算出部５は、活性度の計算に用いるニューラルネットワークの活性化関数Ｆ（ａｃｔｉｖａｔｉｏｎ　ｆｕｎｃｔｉｏｎ）として、線形関数、シグモイド関数、ソフトマックス関数、整流化線形関数（ＲｅＬＵ）などを事前に用意しており、式（３）の加算結果ＡＤＤを活性化関数Ｆの引数とし、下記の式（４）に示すように、第１中間層における“Ｎ”のノードの活性度A_{1M-N}として、その活性化関数の関数値を算出する（ステップＳＴ１５）。

$$A_{1M-N} = F(ADD) \quad (4)$$

ここでは、第１中間層における“Ｎ”のノードの活性度A_{1M-N}の算出例を示したが、第１中間層における他のノードの活性度A_{1M}についても同様に算出する。

【００３１】

第２中間層活性度算出部６は、第１中間層活性度算出部５が第１中間層の各ノードでの活性度A_{1M}を算出すると、第２中間層の各ノードでの活性度A_{2M}を算出する（図４の

ステップＳＴ３）。
第２中間層活性度算出部６による第２中間層の各ノードでの活性度A_{2M}の算出方法は、第１中間層活性度算出部５による第１中間層の各ノードでの活性度A_{1M}の算出方法と同様である。
即ち、第２中間層活性度算出部６は、第２中間層記憶部３に記憶されているインデックスを参照して、第２中間層のノード毎に、当該ノードに接続されている第１中間層の各ノードを確認して、その第１中間層の各ノードでの活性度A_{1M}を取得する。
また、第２中間層活性度算出部６は、第２中間層記憶部３に記憶されているインデックスを参照して、第２中間層のノード毎に、当該ノードに接続されているエッジを確認して、第２中間層記憶部３からそのエッジの重みｗを取得する。
【００３２】
また、第２中間層活性度算出部６は、第２中間層のノード毎に、第２中間層記憶部３から当該ノードのバイアス値B_{2M}を取得する。
第２中間層活性度算出部６は、第２中間層のノード毎に、第１中間層の各ノードでの活性度A_{1M}、エッジの重みｗ、バイアス値B_{2M}を取得すると、第１中間層活性度算出部５と同様の計算方法で、活性度A_{1M}、エッジの重みｗ、バイアス値B_{2M}を用いて、第２中間層のノード毎の活性度A_{2M}を算出する。
【００３３】
第３中間層活性度算出部７は、第２中間層活性度算出部６が第２中間層の各ノードでの活性度A_{2M}を算出すると、第３中間層の各ノードでの活性度A_{3M}を算出する（ステップＳＴ４）。
第３中間層活性度算出部７による第３中間層の各ノードでの活性度A_{3M}の算出方法は、第１中間層活性度算出部５による第１中間層の各ノードでの活性度A_{1M}の算出方法と同様である。
即ち、第３中間層活性度算出部７は、第３中間層記憶部４に記憶されているインデックスを参照して、第３中間層のノード毎に、当該ノードに接続されている第２中間層の各ノードを確認して、その第２中間層の各ノードでの活性度A_{2M}を取得する。
また、第３中間層活性度算出部７は、第３中間層記憶部４に記憶されているインデックスを参照して、第３中間層のノード毎に、当該ノードに接続されているエッジを確認して、第３中間層記憶部４からそのエッジの重みｗを取得する。
【００３４】
また、第３中間層活性度算出部７は、第３中間層のノード毎に、第３中間層記憶部４から当該ノードのバイアス値B_{3M}を取得する。
第３中間層活性度算出部７は、第３中間層のノード毎に、第２中間層の各ノードでの活性度A_{2M}、エッジの重みｗ、バイアス値B_{3M}を取得すると、第１中間層活性度算出部５と同様の計算方法で、活性度A_{2M}、エッジの重みｗ、バイアス値B_{3M}を用いて、第３中間層のノード毎の活性度A_{3M}を算出する。
【００３５】
出力層活性度算出部９は、第３中間層活性度算出部７が第３中間層の各ノードでの活性度A_{3M}を算出すると、出力層の各ノードでの活性度A_{OUT}を算出する（ステップＳＴ５）。
出力層活性度算出部９による出力層の各ノードでの活性度A_{OUT}の算出方法は、第１中間層活性度算出部５による第１中間層の各ノードでの活性度A_{1M}の算出方法と同様である。
即ち、出力層活性度算出部９は、出力層記憶部８に記憶されているインデックスを参照して、出力層のノード毎に、当該ノードに接続されている第３中間層の各ノードを確認して、その第３中間層の各ノードでの活性度A_{3M}を取得する。
また、出力層活性度算出部９は、出力層記憶部８に記憶されているインデックスを参照して、出力層のノード毎に、当該ノードに接続されているエッジを確認して、出力層記憶部８からそのエッジの重みｗを取得する。

【００３６】
また、出力層活性度算出部９は、出力層のノード毎に、出力層記憶部８から当該ノードのバイアス値B_{OUT}を取得する。
出力層活性度算出部９は、出力層のノード毎に、第３中間層の各ノードでの活性度A_{3M}、エッジの重みw、バイアス値B_{OUT}を取得すると、第１中間層活性度算出部５と同様の計算方法で、活性度A_{3M}、エッジの重みw、バイアス値B_{OUT}を用いて、出力層のノード毎の活性度A_{OUT}を算出する。
【００３７】
出力層活性度算出部９により算出された出力層のノード毎の活性度A_{OUT}は、推論装置の推論結果として出力される。
例えば、画像に映っているものが人、犬、猫、自動車のいずれであるかを識別する場合、出力層は、４つのノードから構成され、各ノードの活性度が、それぞれ人、犬、猫、自動車である可能性を示す値になるように学習される。
推論時は、出力層の中で一番活性度が大きいノードを選び、例えば、それが猫である可能性を出力するノードであれば、猫という推論結果を出力する。単なる識別結果のみでなく、活性度を用いた信頼度の算出や回帰予測値出力等の処理を実施してもよい。
【００３８】
以上で明らかなように、この実施の形態１によれば、第１中間層活性度算出部５が、第１中間層記憶部２に記憶されているインデックスを参照して、入力層活性度算出部１により算出された入力層の各ノードでの活性度と第１中間層記憶部２に記憶されている各エッジの重み及びバイアス値の中から、第１中間層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した入力層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第１中間層の各ノードでの活性度を算出するように構成したので、推論を行う際の計算量及びメモリ量を削減することができる効果を奏する。
即ち、第１中間層活性度算出部５では、第１中間層の各ノードと接続関係がある入力層の各ノードについてだけ計算を行えばよいため、推論を行う際の計算量及びメモリ量を大幅に削減することができる。
【００３９】
また、第２中間層活性度算出部６についても、第２中間層の各ノードと接続関係がある第１中間層の各ノードについてだけ計算を行えばよいため、第１中間層活性度算出部５と同様に、推論を行う際の計算量及びメモリ量を大幅に削減することができる。
また、第３中間層活性度算出部７についても、第３中間層の各ノードと接続関係がある第２中間層の各ノードについてだけ計算を行えばよいため、第１中間層活性度算出部５と同様に、推論を行う際の計算量及びメモリ量を大幅に削減することができる。
さらに、出力層活性度算出部９についても、出力層の各ノードと接続関係がある第３中間層の各ノードについてだけ計算を行えばよいため、第１中間層活性度算出部５と同様に、推論を行う際の計算量及びメモリ量を大幅に削減することができる。
【００４０】
この実施の形態１では、第１中間層活性度算出部５、第２中間層活性度算出部６、第３中間層活性度算出部７及び出力層活性度算出部９が、各ノードでの活性度を算出する際、前段の層の各ノードでの活性度とエッジの重みとの積和演算を実施するものを示したが、前段の層の各ノードでの活性度の最大値又は平均値を求め、その最大値又は平均値を式（３）の加算結果ＡＤＤの代わりに用いてもよい。
【００４１】
実施の形態２．
上記実施の形態１では、推論装置が適用するニューラルネットワークを構成している各層の各ノードが、前段又は後段の層の全てのノードとは接続されておらず、一部のノードと接続されているものを示している。
各層の各ノードが、前段又は後段の層の一部のノードと接続される場合でも、ノード間

の接続形態によっては、ノード間を接続する複数のエッジによってループが形成されることがある。

ここで、ニューラルネットワークにおいて、或るノードから同じエッジを一度しか通らないように辿って元のノードに戻る経路をループと称し、ループを構成するエッジの本数をループの長さと称する。

【００４２】

図８は複数のエッジによって形成されるループの例を示す説明図である。

図８（ａ）（ｂ）は４本のエッジによって形成されるループの例を示し、図８（ｃ）は６本のエッジによって形成されるループの例を示し、図８（ｄ）は８本のエッジによって形成されるループの例を示している。

例えば、階層型ニューラルネットワークでは、最短で長さ４のループが生じ得るが、特に、長さ４のループは学習時の逆誤差伝播法で伝播する勾配計算情報が容易に循環するため、推論精度低下の要因になる。また、ベイジアンネットの確率伝播法のように双方向に情報伝播することで推論するモデルでも、短いループの存在によって伝播情報が循環してしまって推論精度低下の要因となる。

【００４３】

そこで、この実施の形態２では、ニューラルネットワークを構成している各層のノード間を接続しているエッジ、即ち、第１中間層のノードと入力層のノードとを接続しているエッジ、第２中間層のノードと第１中間層のノードとを接続しているエッジ、第３中間層のノードと第２中間層のノードとを接続しているエッジ、出力層のノードと第３中間層のノードとを接続しているエッジがループを形成する場合、推論装置が適用するニューラルネットワークは、ループが６本以上のエッジで形成されているニューラルネットワークに限るものとする。

したがって、この実施の形態２では、図８（ａ）（ｂ）に示すように４本のエッジによってループが形成されているニューラルネットワークは、推論装置の適用対象とならないが、図８（ｃ）（ｄ）に示すように６本又は８本のエッジによってループが形成されているニューラルネットワークは、推論装置の適用対象となる。

これにより、ニューラルネットワーク内に存在する長さの短いループに伴う推論精度の低下を抑制することができる効果を奏する。即ち、推論精度を維持したまま計算量及びメモリ量を削減することができる。

【００４４】

実施の形態３．

上記実施の形態１では、推論装置が適用するニューラルネットワークを構成している各層の各ノードが、前段又は後段の層の全てのノードとは接続されておらず、一部のノードと接続されているものを示している。

このとき、第１中間層の各ノードは、入力層における全ノードのうち、ランダムに選択された一部のノードと接続されているものであってもよい。

同様に、第２中間層の各ノードは、第１中間層における全ノードのうち、ランダムに選択された一部のノードと接続されているものであってもよく、また、第３中間層の各ノードは、第２中間層における全ノードのうち、ランダムに選択された一部のノードと接続されているものであってもよい。

また、同様に、出力層の各ノードは、第３中間層における全ノードのうち、ランダムに選択された一部のノードと接続されているものであってもよい。

【００４５】

ランダムに選択される条件として、各層（出力層、第３中間層、第２中間層、第１中間層）の各ノード当りの前段の層（第３中間層、第２中間層、第１中間層、入力層）のノードとの平均接続本数が５０本以下であるという条件を課してもよい。

あるいは、各層（出力層、第３中間層、第２中間層、第１中間層）の各ノード当りの前段の層（第３中間層、第２中間層、第１中間層、入力層）のノードとの平均接続本数が前段の層のノードの個数の１０分の１以下であるという条件を課してもよい。

【００４６】
各層の各ノードが前段の層における全ノードと接続される形態では、各層のノード数がＭで、前段の層のノード数がＮである場合、各層での活性度の計算量とメモリ量がＮ×Ｍのオーダーになるが、各層の各ノード当りの前段の層のノードの平均接続本数ｎ（ｎ＜Ｎ）が、５０本以下であるという条件、あるいは、前段の層のノードの個数Ｎの１０分の１以下であるという条件を課することで、長さの短いループの発生確率を低下させて、推論精度の低下を抑制することができるとともに、計算量及びメモリ量を削減することができる。
【００４７】
実施の形態４．
上記実施の形態１では、推論装置が適用するニューラルネットワークを構成している各層の各ノードが、前段又は後段の層の全てのノードとは接続されておらず、一部のノードと接続されているものを示している。
このとき、第１中間層の各ノードは、入力層における全ノードのうち、隣接していない一部のノードと接続されているようにしてもよい。
同様に、第２中間層の各ノードは、第１中間層における全ノードのうち、隣接していない一部のノードと接続されているようにしてもよく、また、第３中間層の各ノードは、第２中間層における全ノードのうち、隣接していない一部のノードと接続されているようにしてもよい。
また、同様に、出力層の各ノードは、第３中間層における全ノードのうち、隣接していない一部のノードと接続されているようにしてもよい。
例えば、第１中間層における“Ｎ”のノードの場合、入力層における“０”のノードと、“３”のノードとは隣接していないため、第１中間層における“Ｎ”のノードは、入力層における“０”，“３”のノードと接続される形態が許容されるが、入力層における“０”のノードと、“１”のノードとは隣接しているため、第１中間層における“Ｎ”のノードは、入力層における“０”，“１”のノードと接続される形態が許容されない。
【００４８】
この実施の形態４において、接続される形態が許容される条件として、各層（出力層、第３中間層、第２中間層、第１中間層）の各ノード当りの前段の層（第３中間層、第２中間層、第１中間層、入力層）のノードとの平均接続本数が５０本以下であるという条件を課してもよい。
あるいは、各層（出力層、第３中間層、第２中間層、第１中間層）の各ノード当りの前段の層（第３中間層、第２中間層、第１中間層、入力層）のノードとの平均接続本数が前段の層のノードの個数の１０分の１以下であるという条件を課してもよい。
上記の条件を課することで、長さの短いループの発生確率を低下させて、推論精度の低下を抑制することができるとともに、計算量及びメモリ量を削減することができる。
【００４９】
実施の形態５．
上記実施の形態１～４では、推論装置が適用するニューラルネットワークとして、３つの中間層をもつ階層型フィードフォワードのニューラルネットワークを例に挙げているが、中間層は３つより多くても少なくてもよい。また、中間層がないロジスティック回帰モデルのような構造であってもよい。
また、層間のノードを全結合する層や、畳み込みニューラルネットワークのような畳み込み層およびプーリング層や、リカレントニューラルネットワークにおけるＬＳＴＭ（長期短期記憶）ブロックなど、ニューラルネットワークの従来手法と組み合わせてもよい。
ここで、畳み込みニューラルネットワークは、畳み込み層とプーリング層が繰り返された構造になっている。例えば、畳み込み層は、画像の局所的な特徴抽出を担う層であり、プーリング層は、局所毎に特徴をまとめあげる層である。
【００５０】
上記実施の形態１～４では、推論装置が適用するニューラルネットワークとして、階層

型フィードフォワードのニューラルネットワークを例に挙げているが、層を飛ばした接続があってもよいし、同じ層に属するノード同士で接続があってもよいし、接続先と接続元が同一の自己接続があってもよいし、エッジがループを形成するような循環接続があってもよい（リカレントニューラルネットワーク）。

また、自己組織化マップ（ＳＯＭ）、連想記憶モデル、ホップフィールドネットワーク、ボルツマンマシンなど、他のグラフを用いて推論するニューラルネットワークであってもよい。さらに、ニューラルネットワークに限らず、ベイジアンネットワークなど、他のグラフを用いて推論するモデルでもよい。

【００５１】

上記実施の形態１～４では、入力層のノードが０，１，…，Ｎ－１、第１中間層のノードがＮ，Ｎ＋１，…，Ｎ＋Ｍ－１といったように一次元のインデックスを付けているが、入力層のノードが（０，０）、（０，１）、…、（０，Ｎ－１）、第１中間層のノードが（１，０）、（１，１）、…、（１，Ｍ－１）といったように二次元のインデックスを付けてもよいし、メモリのアドレスをインデックスとして用いてもよいし、他のインデックスを付けてもよい。

【００５２】

上記実施の形態１～４では、推論装置が適用するニューラルネットワークとして、エッジの数とエッジ重みの数が一致する例を挙げているが、畳み込みネットワークにおける畳み込みフィルタ係数のように、複数のエッジ重みを共有化するようにしてもよい。

上記実施の形態１～４では、各ノードにおける活性度の計算過程を順に記載しているが、互いに依存しない計算を複数のＣＰＵやＧＰＵを用いて並列化し、さらに高速化することも可能である。

【００５３】

上記実施の形態１～４では、画像データを入力して、画像を分類する画像分類システムを例に挙げているが、データと対応する教師信号が準備できており、教師あり学習を行うことができるならば、データの入力に対して何らかの推論結果を出力する推論システム全般に適用可能である。

例えば、画像を入力して検知したい物体領域の位置や大きさを出力してもよいし、画像を入力して、その画像を説明するテキストを出力してもよいし、ノイズが入った画像を入力して、ノイズを除去した画像を出力してもよいし、画像とテキストを入力して、画像をテキストに従って変換してもよい。

また、音声を入力して音素や単語を出力してもよいし、音声を入力して次に発話される単語を予測してもよいし、音声を入力して、それに対する適切な応答音声を出力してもよいし、テキストを入力して別の言語のテキストを出力してもよいし、時系列を入力して将来の時系列を予測してもよいし、時系列を入力して時系列の状態を推定してもよい。

【００５４】

上記実施の形態１～４では、データと対応する教師信号を用いる教師あり学習によって学習したモデルで推論するシステム例を挙げているが、教師信号のないデータを用いる教師なし学習や半教師あり学習によって学習したモデルで推論するシステムでもよい。

上記実施の形態１～４では、推論装置が、図示せぬデータ入力装置から画像データを与えられて、第１中間層の各ノードでの活性度を算出する例を示したが、図示せぬデータ入力装置が、第１中間層の各ノードでの活性度を算出し、推論装置が、第２及び第３中間層及び出力層の各ノードでの活性度を算出するようにしてもよい。データ入力装置の出力の次元数が入力の次元数よりも少ない場合、データ入力装置がデータ圧縮の機能も併せ持つことになる。

上記実施の形態１～４では、各ノードに対して一度だけ活性度を算出する例を挙げているが、ベイジアンネットワークの確率伝播法のようにノード間で繰り返し何度も情報交換して推論精度を向上させてもよい。

【００５５】

実施の形態６．

上記実施の形態１～４では、推論装置が適用するニューラルネットワークとして、入力層を除く全ての層で枝接続のインデックスを保持している例を挙げているが、一部の層だけが枝接続のインデックスを保持し、他の層では通常のニューラルネットワークと同様の枝接続であるものでもあってもよい。

ここで、枝接続のインデックスとは、図７に示すようなインデックスであり、エッジ重みやバイアス値を含む概念である。

また、通常のニューラルネットワークと同様の枝接続とは、接続先の層における全てのノードと接続されてる枝接続（全接続層の枝接続）を意味するほか、接続先の層における或るノード及び当該ノードの周辺ノードと接続されている畳み込み層やプーリング層などの公知のニューラルネットワークの枝接続を意味する。

【００５６】

図９は入力層に与えられる画像データを１０種のクラスに識別するニューラルネットワークの一例を示す説明図である。

図９は例では、入力層と出力層の間に５つの中間層、即ち、第１中間層、第２中間層、第３中間層、第４中間層及び第５中間層が接続されている。

また、図９は例では、入力層から第１中間層が畳み込み層３１、第１中間層から第２中間層がプーリング層３２、第２中間層から第３中間層が畳み込み層３３、第３中間層から第４中間層がプーリング層３４、第４中間層から第５中間層が上記実施の形態１～４で示しているインデックスを保持する層３５、第５中間層から出力層が全接続層３６である。

このため、第５中間層における各ノードは、図７に示している第１中間層と同様に、第４中間層における接続元のノードを示すインデックスと、その接続に対応するエッジ重み及びバイアス値を保持している。

【００５７】

例えば、入力層に与えられる画像データが、縦６０×横６０画素の画像データであれば、図９のニューラルネットワークでは、３６００個（＝６０×６０×１個）のノードを有する入力層が必要となる。

このとき、例えば、入力層から第１中間層への畳み込み層３１のフィルタサイズが５×５×１、この畳み込み層３１でのマップ数が１００、第１中間層から第２中間層へのプーリング層３２及び第３中間層から第４中間層へのプーリング層３４がフィルタサイズ２×２×１の最大値プーリングである場合、第１中間層のサイズが５６×５６×１００（＝（６０－５＋１）×（６０－５＋１）×１００）、第２中間層のサイズが２８×２８×１００（＝（５６／２）×（５６／２）×１００）となる。

また、第３中間層のサイズが２４×２４×２００（＝（２８－５＋１）×（２８－５＋１）×２００、第４中間層のサイズが１２×１２×２００（＝（２４／２）×（２４／２）×２００）、第５中間層のサイズが１×１×１０００、出力層のノード数が１×１×１０となる。

【００５８】

なお、入力層から第１中間層に情報を伝播する際に伝播値を計算する活性化関数、第２中間層から第３中間層に情報を伝播する際に伝播値を計算する活性化関数や、第４中間層から第５中間層に情報を伝播する際に伝播値を計算する活性化関数として、例えば、ＲｅＬＵ（Ｒｅｃｔｉｆｉｅｄ　Ｌｉｎｅａｒ　Ｕｎｉｔ）が用いられ、第５中間層から出力層に情報を伝播する際に伝播値を計算する活性化関数として、例えば、正規化指数関数であるソフトマックス関数（Ｓｏｆｔｍａｘ関数）が用いられる。

【００５９】

図９のニューラルネットワークでは、入力層から第４中間層までの畳み込み層３１，３３とプーリング層３２，３４によって、入力された画像の位置変化に対してロバストに画像データの特徴量を抽出することができる。

また、第４中間層から第５中間層へのインデックスを保持する層によって、上記実施の形態１～４と同様に、推論を行う際の計算量及びメモリ量を大幅に削減することができる。

【００６０】
この実施の形態６では、画像データが入力層に与えられる例を示しているが、入力層に与えられるデータは画像データに限るものではなく、例えば、センサにより観測されたデータであるセンサ信号、音声やテキストなどのデータなどであってもよい。
また、この実施の形態６では、入力層に与えられる画像データを１０種のクラスに識別する例を示しているが、ニューラルネットワークを構成している出力層を変更することで、画像データのクラスを識別する推論以外の推論を行うようにしてもよい。
例えば、画像データのノイズを取り除くデノイジング、回帰予測や尤度算出などの推論を行うようにしてもよい。
また、推論の目的に合わせて、各層のノード数やフィルタサイズを変えてもよい。
【００６１】
図９のニューラルネットワークでは、畳み込み層３１、プーリング層３２、畳み込み層３３、プーリング層３４、インデックスを保持する層３５、全接続層３６の順序で枝接続している例を示しているが、上記実施の形態１～４に示すインデックスを保持する層が１つでも接続されていればよく、上記の順序以外の順序で枝接続しているものであってもよい。また、プーリング層３２，３４が接続されていないものであってもよい。
【００６２】
実施の形態７．
実施の形態１～６では、推論装置が適用しているニューラルネットワークが有向ループをなしていないフィードフォワードニューラルネットワーク（ＦＦＮＮ）である例を挙げているが、推論装置が、ネットワークの一部が有向ループをなしているリカレントニューラルネットワーク（ＲＮＮ）を適用し、そのリカレントニューラルネットワークの一部の層が、上記実施の形態１～４に示すインデックスを保持する層であるものであってもよい。
【００６３】
図１０は入力層から第１中間層、第１中間層から第２中間層（文脈層）、第２中間層（文脈層）から第１中間層、第１中間層から出力層に枝接続を持っているリカレントニューラルネットワーク（ＲＮＮ）であるエルマンネットワーク（Ｅｌｍａｎ　Ｎｅｔｗｏｒｋ）を示す説明図である。
図１０のエルマンネットワークでは、第２中間層（文脈層）のノード数が、第１中間層のノード数と等しいものとする。
【００６４】
図１１はこの発明の実施の形態７による推論装置を示す構成図であり、図１１において、図１と同一符号は同一または相当部分を示すので説明を省略する。
中間層活性度算出部である第２中間層活性度算出部４１は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出回路１３で実現されるものであり、第１中間層の各ノードの活性度を第２中間層（文脈層）の各ノードの活性度としてコピーする処理を実施する。
この実施の形態７では、第２中間層活性度算出部４１が、第１中間層の各ノードの活性度を第２中間層（文脈層）の各ノードの活性度としてコピーする例を想定しているが、これは一例に過ぎず、例えば、入力層活性度算出部１と同様に、式（１）によって第２中間層（文脈層）の各ノードの活性度を算出するようにしてもよい。
【００６５】
中間層記憶部である第１中間層記憶部４２は例えばＲＡＭやハードディスクなどの記憶媒体からなる中間層記憶装置１２で実現されるものであり、第１中間層の各ノードと第２中間層（文脈層）の各ノードとの接続関係を示すインデックスと、第１中間層のノードと第２中間層のノードとを接続している各エッジの重みと、第１中間層の各ノードに与えられているバイアス値とを記憶している。
中間層活性度算出部である第１中間層活性度算出部４３は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出

回路１３で実現されるものであり、第１中間層記憶部４２に記憶されているインデックスを参照して、第２中間層活性度算出部４１により求められた第２中間層（文脈層）の各ノードでの活性度と第１中間層記憶部４２に記憶されている各エッジの重み及びバイアス値の中から、第１中間層の各ノードと接続関係がある第２中間層（文脈層）の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した第２中間層（文脈層）の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第１中間層の各ノードでの活性度を算出する処理を実施する。

【００６６】

図１１では、推論装置の構成要素である入力層活性度算出部１、第１中間層記憶部２、第１中間層活性度算出部５、第２中間層活性度算出部４１、第１中間層記憶部４２、第１中間層活性度算出部４３、出力層記憶部８及び出力層活性度算出部９のそれぞれが専用のハードウェアで構成されているものを想定しているが、推論装置がコンピュータで構成されていてもよい。

推論装置がコンピュータで構成される場合、第１中間層記憶部２、第１中間層記憶部４２及び出力層記憶部８を図３に示すコンピュータのメモリ２１上に構成するとともに、入力層活性度算出部１、第１中間層活性度算出部５、第２中間層活性度算出部４１、第１中間層活性度算出部４３及び出力層活性度算出部９の処理内容を記述しているプログラムを図３に示すコンピュータのメモリ２１に格納し、当該コンピュータのプロセッサ２２がメモリ２１に格納されているプログラムを実行するようにすればよい。

【００６７】

次に動作について説明する。

ただし、第２中間層活性度算出部４１、第１中間層記憶部４２及び第１中間層活性度算出部４３以外は、上記実施の形態１と同様であるため、ここでは、第２中間層活性度算出部４１、第１中間層記憶部４２及び第１中間層活性度算出部４３についてのみ説明する。

第２中間層活性度算出部４１は、第１中間層活性度算出部５が上記実施の形態１と同様に、入力層活性度算出部１により算出された入力層の各ノードの活性度を用いて、第１中間層の各ノードの活性度A_{1M}を算出すると、第１中間層の各ノードの活性度A_{1M}を第２中間層（文脈層）の各ノードの活性度A_{2M}としてコピーする。

これにより、ある時刻ｔにおける第２中間層の各ノードの活性度A_{2M}は、時刻ｔにおける第１中間層の各ノードの活性度A_{1M}と同じになる。

【００６８】

第１中間層活性度算出部４３は、第２中間層活性度算出部４１が第２中間層（文脈層）の各ノードの活性度A_{2M}を求めると、第１中間層の各ノードの活性度A'_{1M}を算出する。

第１中間層活性度算出部４３による第１中間層の各ノードでの活性度A'_{1M}の算出方法は、第１中間層活性度算出部５による第１中間層の各ノードでの活性度A_{1M}の算出方法と同様である。

即ち、第１中間層活性度算出部４３は、第１中間層記憶部４２に記憶されているインデックスを参照して、第１中間層のノード毎に、当該ノードに接続されている第２中間層（文脈層）の各ノードを確認して、第２中間層（文脈層）の各ノードでの活性度A_{2M}を取得する。

また、第１中間層活性度算出部４３は、第１中間層記憶部４２に記憶されているインデックスを参照して、第１中間層のノード毎に、当該ノードに接続されているエッジ（第２中間層のノードと接続されているエッジ）を確認して、第１中間層記憶部４２からそのエッジの重みｗを取得する。

【００６９】

また、第１中間層活性度算出部４３は、第１中間層のノード毎に、第１中間層記憶部４２から当該ノードのバイアス値B_{1M}を取得する。

第１中間層活性度算出部４３は、第１中間層のノード毎に、第２中間層（文脈層）の各ノードでの活性度A_{2M}、エッジの重みｗ、バイアス値B_{1M}を取得すると、第１中間層

活性度算出部５と同様の計算方法で、活性度A_{2M}、エッジの重みw、バイアス値B_{1M}を用いて、第１中間層のノード毎の活性度A'_{1M}を算出する。

出力層活性度算出部９は、第１中間層活性度算出部４３が第１中間層の各ノードでの活性度A'_{1M}を算出すると、第１中間層の各ノードでの活性度A'_{1M}を用いて、出力層の各ノードでの活性度A_{OUT}を算出する。

出力層活性度算出部９による出力層の各ノードでの活性度A_{OUT}の算出方法は上記実施の形態１と同様である。

【００７０】

以上で明らかなように、この実施の形態７によれば、第１中間層活性度算出部４３が、第１中間層記憶部４２に記憶されているインデックスを参照して、第２中間層活性度算出部４１により求められた第２中間層（文脈層）の各ノードでの活性度と第１中間層記憶部４２に記憶されている各エッジの重み及びバイアス値の中から、第１中間層の各ノードと接続関係がある第２中間層（文脈層）の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した第２中間層（文脈層）の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第１中間層の各ノードでの活性度を算出するように構成したので、ネットワークの一部が有向ループをなしているリカレントニューラルネットワーク（ＲＮＮ）を適用する場合であっても、推論を行う際の計算量及びメモリ量を削減することができる効果を奏する。

即ち、第１中間層活性度算出部４３では、第１中間層の各ノードと接続関係がある第２中間層（文脈層）の各ノードについてだけ計算を行えばよいため、推論を行う際の計算量及びメモリ量を大幅に削減することができる。

【００７１】

この実施の形態７では、推論装置が適用するリカレントニューラルネットワーク（ＲＮＮ）がエルマンネットワークである例に挙げているが、これは一例に過ぎず、例えば、Ｊｏｒｄａｎ　Ｎｅｔｗｏｒｋや、ＬＳＴＭ（Ｌｏｎｇ　Ｓｈｏｒｔ　Ｔｅｒｍ　Ｍｅｍｏｒｙ）ブロックを持つリカレントニューラルネットワーク、階層型のリカレントニューラルネットワーク、双方向のリカレントニューラルネットワーク、連続時間のリカレントニューラルネットワークなどを適用するものであってもよい。

ただし、この実施の形態７では、いずれのリカレントニューラルネットワークを適用する場合でも、一部の層で枝接続のインデックスを保持するものとする。

【００７２】

実施の形態８.

上記実施の形態１～７では、推論装置が適用するニューラルネットワークとして、同じ層のノード同士の枝接続や自己接続が無いフィードフォワードニューラルネットワーク（ＦＦＮＮ）やリカレントニューラルネットワーク（ＲＮＮ）である例に挙げているが、同じ層のノード同士の枝接続や、接続元ノードと接続先ノードが同一のノードである自己接続が有るフィードフォワードニューラルネットワーク（ＦＦＮＮ）やリカレントニューラルネットワーク（ＲＮＮ）を適用するようにしてもよい。また、層を飛ばした接続があるフィードフォワードニューラルネットワーク（ＦＦＮＮ）やリカレントニューラルネットワーク（ＲＮＮ）を適用するようにしてもよい。

【００７３】

図１２は中間層内でノード同士の枝接続や自己接続が存在するとともに、入力層から中間層を飛ばして出力層へ接続する枝が存在するニューラルネットワークであるＥｃｈｏ　Ｓｔａｔｅ　Ｎｅｔｗｏｒｋの例を示す説明図である。

図１２において、中間層のノード同士の枝接続や自己接続は、中間層から中間層への枝接続とみなすことができるため、Ｅｃｈｏ　Ｓｔａｔｅ　Ｎｅｔｗｏｒｋは、層単位では図１３のように表すことができる。

【００７４】

図１４はこの発明の実施の形態８による推論装置を示す構成図であり、図１４において、図１と同一符号は同一または相当部分を示すので説明を省略する。

中間層記憶部５１は例えばＲＡＭやハードディスクなどの記憶媒体からなる中間層記憶装置１２で実現されるものであり、中間層の各ノードと入力層又は出力層の各ノードとの接続関係を示すインデックスと、中間層のノードと入力層又は出力層のノードとを接続している各エッジの重みと、中間層の各ノードに与えられているバイアス値とを記憶している。

また、中間層記憶部５１は中間層のノード同士の枝接続や自己接続の関係を示すインデックスと、中間層のノード同士の枝接続や自己接続している各エッジの重みと、中間層の各ノードに与えられているバイアス値とを記憶している。

【００７５】

中間層活性度算出部５２は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている中間層活性度算出回路１３で実現されるものであり、図１の第１中間層活性度算出部５と同様に、中間層記憶部５１に記憶されているインデックスを参照して、入力層活性度算出部１により算出された入力層の各ノードでの活性度又は出力層活性度算出部５４により算出された出力層の各ノードでの活性度と中間層記憶部５１に記憶されている各エッジの重み及びバイアス値の中から、中間層の各ノードと接続関係がある入力層又は出力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した入力層又は出力層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、中間層の各ノードでの活性度を算出する処理を実施する。

また、中間層活性度算出部５２は中間層記憶部５１に記憶されているインデックスを参照して、中間層における算出済みの各ノードでの活性度と中間層記憶部５１に記憶されている各エッジの重み及びバイアス値の中から、中間層における接続先の各ノードと接続関係がある中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、中間層における接続先の各ノードでの活性度を算出する処理を実施する。

ここで、中間層における接続先のノードとは、中間層における他のノードと接続されているノード、あるいは、中間層において、自己のノードと接続されているノードを意味する。

また、中間層における接続元のノードとは、接続先のノードと接続されている中間層における他のノード、あるいは、中間層における自己接続のノードを意味する。

【００７６】

出力層記憶部５３は例えばＲＡＭやハードディスクなどの記憶媒体からなる出力層記憶装置１４で実現されるものであり、出力層の各ノードと入力層又は中間層の各ノードとの接続関係を示すインデックス（接続情報）を記憶している。

また、出力層記憶部５３は出力層のノードが入力層のノードと接続されていれば、出力層のノードと入力層のノードとを接続している各エッジの重みと、入力層のノードと接続されている出力層のノードに与えられているバイアス値とを記憶している。

また、出力層記憶部５３は出力層のノードが中間層のノードと接続されていれば、出力層のノードと中間層のノードとを接続している各エッジの重みと、中間層のノードと接続されている出力層のノードに与えられているバイアス値とを記憶している。

【００７７】

出力層活性度算出部５４は例えばＣＰＵを実装している半導体集積回路、あるいは、ワンチップマイコンなどで構成されている出力層活性度算出回路１５で実現されるものであり、出力層のノードと接続されているノードが入力層のノードであれば、入力層活性度算出部１により算出された入力層の各ノードでの活性度と出力層記憶部５３に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、出力層のノードと接続されているノードが中間層のノードであれば、中間層活性度算出部５２により算出された中間層のノードでの活性度と出力層記憶部５３に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある中間層のノードでの活性度と各エッジ

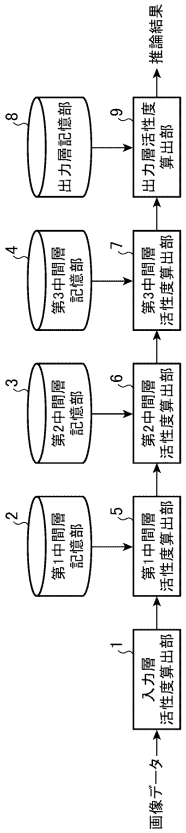
の重みとバイアス値とを取得する処理を実施する。
また、出力層活性度算出部５４は取得した入力層又は中間層のノードでの活性度と各エッジの重みとバイアス値とを用いて、出力層の各ノードでの活性度を算出する処理を実施する。
【００７８】
図１４では、推論装置の構成要素である入力層活性度算出部１、中間層記憶部５１、中間層活性度算出部５２、出力層記憶部５３及び出力層活性度算出部５４のそれぞれが専用のハードウェアで構成されているものを想定しているが、推論装置がコンピュータで構成されていてもよい。
推論装置がコンピュータで構成される場合、中間層記憶部５１及び出力層記憶部５３を図３に示すコンピュータのメモリ２１上に構成するとともに、入力層活性度算出部１、中間層活性度算出部５２及び出力層活性度算出部５４の処理内容を記述しているプログラムを図３に示すコンピュータのメモリ２１に格納し、当該コンピュータのプロセッサ２２がメモリ２１に格納されているプログラムを実行するようにすればよい。
【００７９】
次に動作について説明する。
中間層活性度算出部５２は、入力層活性度算出部１が上記実施の形態１と同様に入力層の各ノードの活性度を算出すると、中間層記憶部５１に記憶されているインデックスを参照して、中間層の各ノードのうち、入力層のノードと接続関係があるノードを確認するとともに、出力層のノードと接続関係があるノードを確認する。
中間層活性度算出部５２は、中間層の各ノードのうち、入力層のノードと接続関係があるノードの活性度については、図１の第１中間層活性度算出部５と同様に、入力層の各ノードの活性度を用いて算出する。
また、中間層の各ノードのうち、出力層のノードと接続関係があるノードの活性度については、出力層活性度算出部５４により算出された出力層の活性度を用いて算出する。
中間層活性度算出部５２による出力層のノードと接続関係があるノードの活性度の算出方法は、入力層のノードと接続関係があるノードの活性度の算出方法と同様であり、活性度算出対象のノードが接続されているノードが、入力層のノードではなく、出力層のノードである点だけが相違している。
【００８０】
中間層活性度算出部５２は、入力層又は出力層のノードと接続関係がある中間層のノードの活性度を算出すると、中間層記憶部５１に記憶されているインデックスを参照して、中間層の各ノードのうち、接続先のノード（中間層における他のノードと接続されているノード、あるいは、中間層において、自己のノードと接続されているノード）と接続関係がある接続元のノードを確認する。
中間層活性度算出部５２は、接続先のノードと接続関係がある接続元のノードを確認すると、中間層における算出済みの各ノードでの活性度と中間層記憶部５１に記憶されている各エッジの重み及びバイアス値の中から、中間層における接続先の各ノードと接続関係がある中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを取得する。
中間層における接続元のノードが、入力層又は出力層のノードと接続関係がある中間層のノードであれば、先に説明したように既に算出済みである。このため、入力層又は出力層のノードと接続関係がある中間層のノードに近いノードから順番に、活性度算出対象のノード（接続先のノード）とすればよい。
【００８１】
中間層活性度算出部５２は、その取得した中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、中間層における接続先の各ノードでの活性度を算出する。
中間層活性度算出部５２による中間層における接続先のノードでの活性度の算出方法は、入力層のノードと接続関係があるノードの活性度の算出方法と同様であり、接続元のノ

ードが、入力層のノードではなく、中間層のノードである点だけが相違している。
【００８２】
出力層活性度算出部５４は、出力層記憶部５３に記憶されているインデックスを参照して、出力層の各ノードが接続されている入力層又は中間層のノードを確認する。
出力層活性度算出部５４は、出力層のノードと接続されているノードが入力層のノードであれば、入力層活性度算出部１により算出された入力層の各ノードでの活性度と出力層記憶部５３に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得する。一方、出力層のノードと接続されているノードが中間層のノードであれば、中間層活性度算出部５２により算出された中間層のノードでの活性度と出力層記憶部５３に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある中間層のノードでの活性度と各エッジの重みとバイアス値とを取得する。
出力層活性度算出部５４は、入力層又は中間層のノードでの活性度と各エッジの重みとバイアス値とを取得すると、その取得した入力層又は中間層のノードでの活性度と各エッジの重みとバイアス値とを用いて、出力層の各ノードでの活性度を算出する。
【００８３】
以上で明らかなように、この実施の形態８によれば、中間層活性度算出部５２が、中間層記憶部５１に記憶されているインデックスを参照して、中間層における算出済みの各ノードでの活性度と中間層記憶部５１に記憶されている各エッジの重み及びバイアス値の中から、中間層における接続先の各ノードと接続関係がある中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した中間層における接続元の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、中間層における接続先の各ノードでの活性度を算出するように構成したので、中間層内でノード同士の枝接続や自己接続が存在するニューラルネットワークを適用する場合であっても、推論を行う際の計算量及びメモリ量を削減することができる効果を奏する。
【００８４】
また、この実施の形態８によれば、出力層活性度算出部５４が、出力層のノードと接続されているノードが入力層のノードであれば、入力層活性度算出部１により算出された入力層の各ノードでの活性度と出力層記憶部５３に記憶されている各エッジの重み及びバイアス値の中から、出力層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した入力層のノードでの活性度と各エッジの重みとバイアス値とを用いて、出力層の各ノードでの活性度を算出するように構成したので、入力層から中間層を飛ばして出力層へ接続する枝が存在するニューラルネットワークを適用する場合であっても、推論を行う際の計算量及びメモリ量を削減することができる効果を奏する。
【００８５】
この実施の形態８では、推論装置が適用するニューラルネットワークがＥｃｈｏ Ｓｔａｔｅ Ｎｅｔｗｏｒｋである例に挙げているが、これは一例に過ぎず、例えば、全結合のリカレントニューラルネットワーク、ホップフィールドネットワーク、ボルツマンマシンなどを適用するものであってもよい。
ただし、この実施の形態８では、いずれのニューラルネットワークを適用する場合でも、一部の層で枝接続のインデックスを保持するものとする。
【００８６】
実施の形態９．
上記実施の形態１～６では、推論装置が適用しているニューラルネットワークがフィードフォワードニューラルネットワーク（ＦＦＮＮ）である例を挙げ、上記実施の形態７，８では、推論装置が適用しているニューラルネットワークがリカレントニューラルネットワーク（ＲＮＮ）である例を挙げているが、これは一例に過ぎず、推論装置が下記に示すようなニューラルネットワークを適用するものであってもよい。ただし、いずれのニューラルネットワークを適用する場合でも、一部の層で枝接続のインデックスを保持するもの

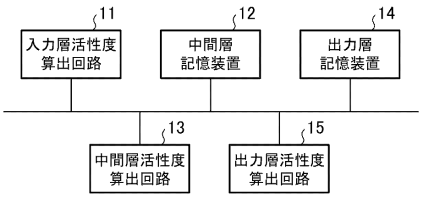
とする。
【００８７】
例えば、推論装置が適用するニューラルネットワークとして、放射基底関数（ＲＢＦ）ネットワーク、自己組織化マップ（ＳＯＭ）、学習ベクトル量子化法（ＬＶＱ）、モジュールニューラルネットワーク、スパイキングニューラルネットワーク、動的ニューラルネットワーク、カスケードニューラルネットワーク、階層型時間記憶（ＨＴＭ）等のＦＦＮＮやＲＮＮ以外のニューラルネットワークが考えられる。
【００８８】
実施の形態１０.
上記実施の形態１～９では、推論装置の学習手法が、データと対応する教師信号を用いる教師あり学習、教師信号のないデータを用いる教師なし学習、あるいは、半教師あり学習である例を挙げているが、推論装置の学習手法が、強化学習であってもよい。
ここで、強化学習とは、ある環境下におけるエージェントが、現在の状態を観測し、取るべき行動を決定するためのモデルを学習する手法である。エージェントは、コンピュータのユーザが連続した操作をしなくても、自律的に情報収集や状況判断を行って適切な処理動作を実行する機能を意味する。
エージェントが行動を選択すると、環境から報酬を得るが、強化学習では、一連の行動で報酬を最大化できるようなポリシーを学習する。
【００８９】
強化学習では、現在の状態、あるいは、行動がどのくらい良いかを計る指標として、状態ｓの価値を表す状態価値関数Ｖ（ｓ）や、状態ｓのときに行動ａを選択することで環境から得られる報酬を表す行動価値関数Ｑ（ｓ，ａ）を用いる。強化学習のアルゴリズムとしては、ＳａｒｓａやＱ学習（Ｑ－ｌｅａｒｎｉｎｇ）等のＴＤ（Ｔｅｍｐｏｒａｌ　Ｄｉｆｆｅｒｅｎｃｅ：時間差分）学習が用いられる。
推論装置の学習手法が強化学習である場合、状態ｓを入力として、状態価値関数Ｖ（ｓ）や行動価値関数Ｑ（ｓ，ａ）を出力するニューラルネットワークを学習し、これらを用いてＴＤ学習することになる。即ち、一部の層で枝接続のインデックスを保持するようしたニューラルネットワークを用いて、状態価値関数Ｖ（ｓ）や行動価値関数Ｑ（ｓ，ａ）を計算して、強化学習を行う。
【００９０】
なお、本願発明はその発明の範囲内において、各実施の形態の自由な組み合わせ、あるいは各実施の形態の任意の構成要素の変形、もしくは各実施の形態において任意の構成要素の省略が可能である。
【産業上の利用可能性】
【００９１】
この発明に係る推論装置は、推論を行う際の計算量やメモリ量を削減する必要が高いものに適している。
【符号の説明】
【００９２】
１　入力層活性度算出部、２　第１中間層記憶部（中間層記憶部）、３　第２中間層記憶部（中間層記憶部）、４　第３中間層記憶部（中間層記憶部）、５　第１中間層活性度算出部（中間層活性度算出部）、６　第２中間層活性度算出部（中間層活性度算出部）、７　第３中間層活性度算出部（中間層活性度算出部）、８　出力層記憶部、９　出力層活性度算出部、１１　入力層活性度算出回路、１２　中間層記憶装置、１３　中間層活性度算出回路、１４　出力層記憶装置、１５　出力層活性度算出回路、２１　メモリ、２２　プロセッサ、３１　畳み込み層、３２　プーリング層、３３　畳み込み層、３４　プーリング層、３５　インデックスを保持する層、３６　全接続層、４１　第２中間層活性度算出部（中間層活性度算出部）、４２　第１中間層記憶部（中間層記憶部）、４３　第１中間層活性度算出部（中間層活性度算出部）、５１　中間層記憶部、５２　中間層活性度算出部、５３　出力層記憶部、５４　出力層活性度算出部。

【要約】
第１中間層活性度算出部（５）が、第１中間層記憶部（２）に記憶されているインデックスを参照して、入力層活性度算出部（１）により算出された入力層の各ノードでの活性度と第１中間層記憶部（２）に記憶されている各エッジの重み及びバイアス値の中から、第１中間層の各ノードと接続関係がある入力層の各ノードでの活性度と各エッジの重みとバイアス値とを取得し、その取得した入力層の各ノードでの活性度と各エッジの重みとバイアス値とを用いて、第１中間層の各ノードでの活性度を算出する。これにより、推論を行う際の計算量及びメモリ量を削減することができる。また、より高い推論精度を得ることができる。

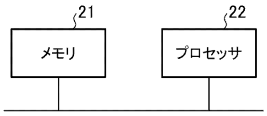
【図１】



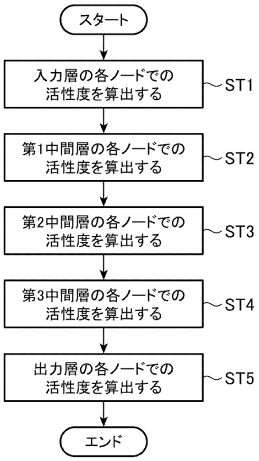
【図２】



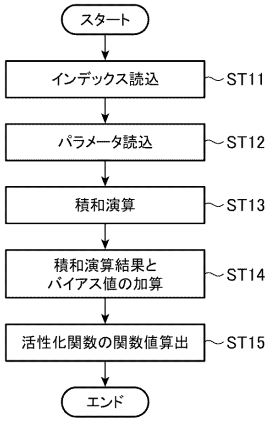
【図３】



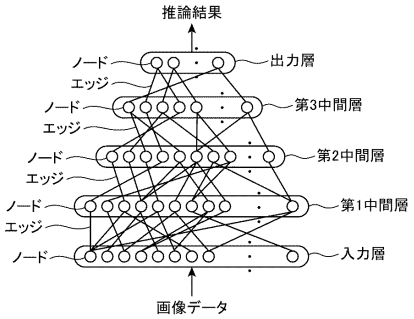
【図４】


スタート
入力層の各ノードでの活性度を算出する
ST1
第1中間層の各ノードでの活性度を算出する
ST2
第2中間層の各ノードでの活性度を算出する
ST3
第3中間層の各ノードでの活性度を算出する
ST4
出力層の各ノードでの活性度を算出する
ST5
エンド

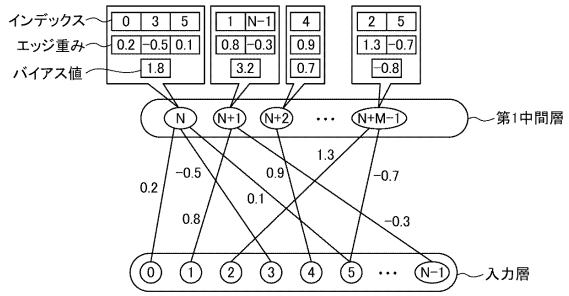
【図５】


スタート
インデックス読込
ST11
パラメータ読込
ST12
積和演算
ST13
積和演算結果とバイアス値の加算
ST14
活性化関数の関数値算出
ST15
エンド

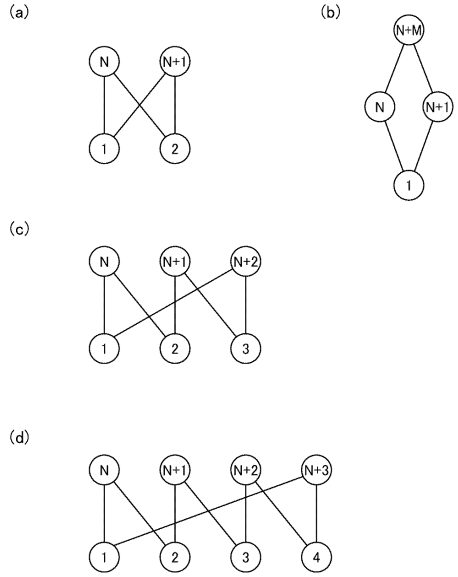
【図６】


推論結果
ノード
エッジ
出力層
第3中間層
第2中間層
第1中間層
入力層
画像データ

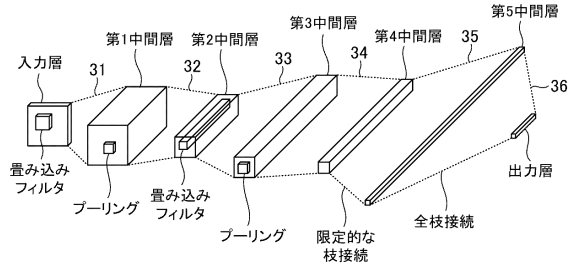
【図７】


インデックス
エッジ重み
バイアス値
第1中間層
入力層
1.3
0.2
−0.5
0.9
−0.7
0.1
0.8
−0.3

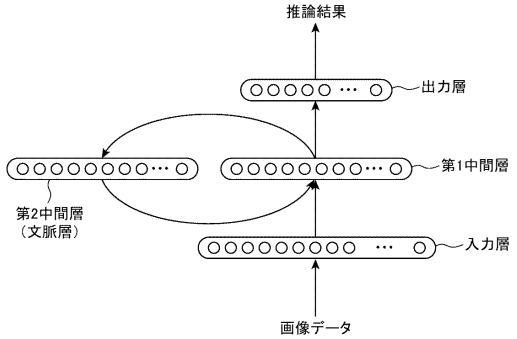
【図８】


(a)
(b)
(c)
(d)

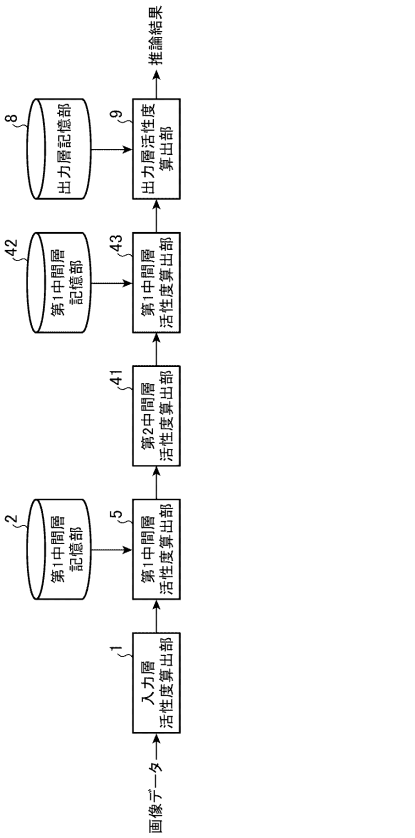
【図９】



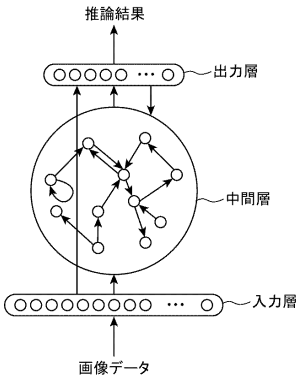
【図１０】



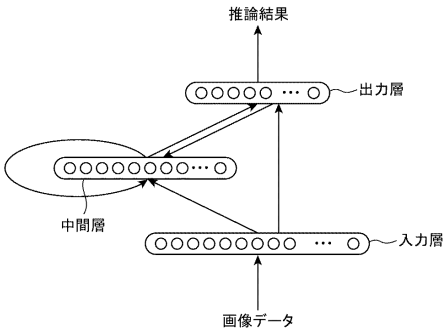
【図１１】



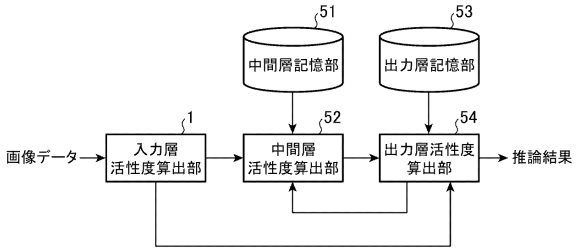
【図１２】



【図１３】



【図１４】



フロントページの続き

(72)発明者　松本　渉
東京都千代田区丸の内二丁目７番３号　三菱電機株式会社内
(72)発明者　吉村　玄太
東京都千代田区丸の内二丁目７番３号　三菱電機株式会社内
(72)発明者　趙　雄心
東京都千代田区丸の内二丁目７番３号　三菱電機株式会社内

審査官　多賀　実

(56)参考文献　特開平３－５５６５８（ＪＰ，Ａ）
特開平２－２３６６５９（ＪＰ，Ａ）
特開２００２－２５１６０１（ＪＰ，Ａ）
特開平４－３５５８８９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ３／０２－３／１０
Ｇ０６Ｔ７／００