

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2019年10月31日(31.10.2019)



(10) 国際公開番号

WO 2019/208564 A1

(51) 国際特許分類:
G06N 3/08 (2006.01)

(21) 国際出願番号: PCT/JP2019/017216

(22) 国際出願日: 2019年4月23日(23.04.2019)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2018-084760 2018年4月26日(26.04.2018) JP

(71) 出願人: 日本電信電話株式会社 (**NIPPON TELEGRAPH AND TELEPHONE CORPORATION**) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).

(72) 発明者: 森谷 崇史 (**MORIYA, Takafumi**); 〒1808585 東京都武蔵野市緑町三丁目9番1号 N T T 知的財産センタ内 Tokyo (JP). 山口 義和 (**YAMAGUCHI, Yoshikazu**); 〒1808585 東京都武蔵野市緑町三丁目9番1号 N T T 知的財産センタ内 Tokyo (JP).

(74) 代理人: 中尾 直樹, 外 (**NAKAO, Naoki et al.**); 〒1600022 東京都新宿区新宿三丁目1番2号 新宿N S Oビル6階 Tokyo (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH,

(54) **Title:** NEURAL NETWORK LEARNING DEVICE, NEURAL NETWORK LEARNING METHOD, AND PROGRAM

(54) 発明の名称: ニューラルネットワーク学習装置、ニューラルネットワーク学習方法、プログラム

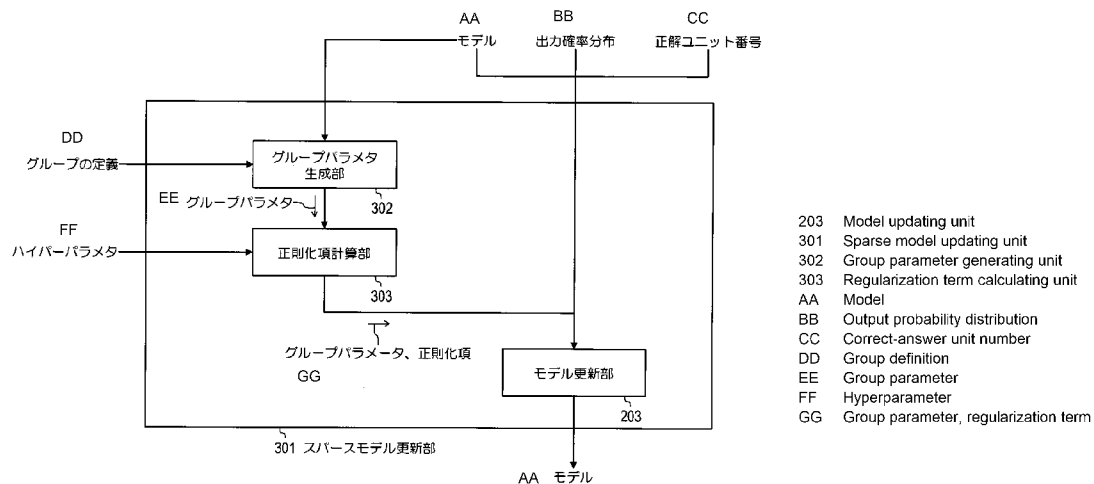


図9

(57) **Abstract:** Provided is a neural network learning device capable of adjusting reduction amount of a model size. The neural network learning device comprises a group parameter generating unit that separates model parameters of a neural network model into arbitrarily defined groups and generates group parameters representing features of the respective groups, a regularization term calculating unit that calculates a regularization term under the assumption that distribution of the group parameters conforms to distribution defined by a hyperparameter as a parameter defining the feature of the distribution, and a model updating unit that calculates loss function from a correct-answer label in teacher data, output probability distribution obtained by inputting feature quantity corresponding to the correct-answer label in the teacher data into the neural network model, and the regularization term and updates the neural network model so as to reduce a

KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類：

- 一 国際調査報告（条約第21条(3)）

value of the loss function.

(57) 要約：モデルサイズの削減量を調整することができるニューラルネットワーク学習装置を提供する。ニューラルネットワークモデルのモデルパラメタを、任意に定義したグループにグループ分けし、各グループの特徴を表すグループパラメタを生成するグループパラメタ生成部と、グループパラメタの分布が、分布の特徴を規定するパラメタであるハイパーパラメタによって規定される分布に従うことを仮定して正則化項を計算する正則化項計算部と、教師データにおける正解ラベルと、教師データにおける正解ラベルに対応する特徴量をニューラルネットワークモデルに入力して得られる出力確率分布と、正則化項から損失関数を計算し、損失関数の値を減少させるようにニューラルネットワークモデルを更新するモデル更新部を含む。

明 細 書

発明の名称：

ニューラルネットワーク学習装置、ニューラルネットワーク学習方法、プログラム

技術分野

[0001] 本発明はニューラルネットワークのモデルパラメタをスパースに学習するニューラルネットワーク学習装置、ニューラルネットワーク学習方法、およびプログラムに関する。

背景技術

[0002] <従来技術 1>

一般的なニューラルネットワーク学習方法の例として非特許文献 1 が開示されている。以下、この学習方法を従来技術 1 と呼称する。非特許文献 1 の“TRAINING DEEP NEURAL NETWORKS”の節には、音声認識用のニューラルネットワーク型音響モデル（以下、単に「音響モデル」または単に「モデル」とも呼称する）を学習する方法が開示されている。

[0003] 従来技術 1 では、事前に学習データの各サンプルから抽出した特徴量（実数ベクトル）と各特徴量に対応する正解ユニット番号（正解ラベル）のペア（教師データ）、および適当な初期モデルを用意する。初期モデルとしては、各パラメタに乱数を割り当てたニューラルネットワークや、既に別の学習データで学習済みのニューラルネットワークなどが利用できる。

[0004] 以下、図 1 を参照して従来技術 1 のニューラルネットワーク学習方法を実行する、ニューラルネットワーク学習装置 100 の構成について説明する。同図に示すように、ニューラルネットワーク学習装置 100 は、中間特徴量抽出部 101 と、出力確率分布計算部 102 と、モデル更新部 103 を含む構成である。以下、図 2 を参照して、各構成要件の動作について説明する。

[0005] [中間特徴量抽出部 101]

入力：特徴量

出力：中間特徴量

処理：

中間特徴量抽出部 101 は、入力された特徴量から、出力確率分布計算部 102 において正解ユニットを識別しやすくするための中間特徴量（非特許文献 1 の式(1)）を抽出する（S101）。この中間特徴量抽出部 101 は複数の層のニューラルネットワークで構築されており、層の数だけ中間特徴量を抽出する計算が行われる。

[0006] [出力確率分布計算部 102]

入力：中間特徴量

出力：出力確率分布

処理：

出力確率分布計算部 102 は、中間特徴量抽出部 101 で抽出した中間特徴量を現在のモデルに入力して出力層の各ユニットの確率を並べた出力確率分布（非特許文献 1 の式(2)）を計算する（S102）。

[0007] この出力確率分布計算部 102 では、音声認識の場合、音声の特徴量を識別しやすくした中間特徴量がどの音声の出力シンボル（音素状態）であるかを計算し、入力した音声の特徴量に対応した出力確率分布を得る。

[0008] [モデル更新部 103]

入力：モデル（更新前）、出力確率分布、正解ユニット番号

出力：モデル（更新後）

処理：

モデル更新部 103 は、正解ユニット番号と出力確率分布計算部 102 より得られる出力確率分布から損失関数 $L(w)=E(w)$ （非特許文献 1 の式(3)）を計算し、損失関数 $L(w)=E(w)$ の値を減少させるように（非特許文献 1 の式(4)によって）モデルを更新する（S103）。

[0009] 更新されるニューラルネットワークモデル内のパラメタ（以下、モデルパラメタと呼称する）は非特許文献 1 の式(1)の重み w とバイアス b である。学習データの特徴量と正解ユニット番号の各ペアに対して、上記の中間特徴量の

抽出→出力確率分布計算→モデル更新の処理を繰り返し、所定回数（通常、数千万～数億回）の繰り返しが完了した時点のモデルを学習済みモデルとして利用する。

[0010] <従来技術 2>

一方、非特許文献 2 には、ニューラルネットワークにおけるモデルサイズを削減しながら学習する方法が示されている。以下、この学習方法を従来技術 2 と呼称する。

[0011] 一般的なニューラルネットワークの学習における損失関数は以下の式で表される。

$$L(w)=E(w)$$

この $E(w)$ は非特許文献 1 の式 (3) の C であり、 w は従来技術 1 において中間特徴量抽出部 101 および出力確率計算部 102 が学習するモデルパラメタである。非特許文献 2 では上式に正則化を加えることでニューラルネットワークのモデルパラメタの一部がスパース (0 に近い値) となるような学習を行う。従来技術 2 では、モデルパラメタの更新部をスパースモデル更新部と呼ぶ。スパースモデル更新部は一般的な損失関数に正則化項を加えた式

$$L(w)=E(w)+\lambda R(w)$$

によりモデル更新を実行する。この式の第 2 項 $\lambda R(w)$ は正則化項であり、非特許文献 2 では Ridge (L2) と Group Lasso と呼ばれる正則化項が用いられる。 λ は正則化項の影響を調整するためのハイパーパラメタである。以下に各層 l における重みパラメタ w のみを更新する場合の L2 ($R_{L2}(w)$) と Group Lasso ($R_{group}(w)$) の正則化項を示す。

[0012] [数 1]

$$R_{L2}(w)=\frac{1}{2}\|w^l\|^2$$

$$R_{group}(w)=\sum_{j=1}^{N_{l-1}}\|w_j^l\|$$

[0013] Group Lasso ではパラメタ間で任意のグルーピングを行うことが可能であり、非特許文献 2 ではグループの単位をニューラルネットの素子（行列 W の行あ

るいは列ごと)としている。 $R_{\text{group}}(w)$ における

[0014] [数2]

$$\sum_{j=1}^{N_{l-1}} \|w_j^l\|$$

[0015] は l 層の1つの素子と $l-1$ 層の全素子 ($j=1, \dots, N_{l-1}$) 間のパラメタである重みの和を表している。

[0016] 正則化項は本来過学習を避ける技術であり、目的によって様々な正則化項が存在する。非特許文献2の式(2)ではGroup LassoやRidge (L2)を用いている。非特許文献2にはGroup Lassoを用いることで自分で決めたグループ(例: 行列における1行ずつをグループとする)ごとにスパースとなるように学習し、学習後のモデルパラメタから利用者が決めた閾値よりも小さい値をもつグループのモデルパラメタを削除することでモデル全体のサイズを削減したことが開示されている。

[0017] 以下、図3を参照して従来技術2のニューラルネットワーク学習方法を実行する、ニューラルネットワーク学習装置200の構成について説明する。同図に示すように、ニューラルネットワーク学習装置200は、中間特徴量抽出部101と、出力確率分布計算部102と、スパースモデル更新部201を含み、中間特徴量抽出部101と、出力確率分布計算部102については、従来技術1の同名の構成要件と同じ動作を実行する。以下、図4を参照して、スパースモデル更新部201の動作について説明する。

[0018] [スパースモデル更新部201]

入力: モデル(更新前)、出力確率分布、正解ユニット番号、ハイパーパラメタ

出力: スパースなモデル(更新後)

処理:

スパースモデル更新部201は、正則化項 $\lambda R(w)$ を計算し、正解ユニット番号と出力確率分布と正則化項 $\lambda R(w)$ から損失関数を計算し、損失関数の値を減少させるようにモデルを更新し、正則化を行わないモデル更新部103で得られるモデルよりもスパースなモデルを出力する(S201)。正則化

項を用いる場合の損失関数を以下に示す。

$$L(w) = E(w) + \lambda R(w)$$

以下、図5を参照して、スパースモデル更新部201の詳細について説明する。同図に示すように、スパースモデル更新部201は、正則化項計算部202と、モデル更新部203を含む構成である。以下、図6を参照して、スパースモデル更新部201内の各構成要件の動作について説明する。

[0019] [正則化項計算部202]

入力：モデル（更新前）、ハイパーパラメタ

出力：正則化項

処理：

正則化項計算部202は、モデルパラメタと、損失関数への影響を調整するためのハイパーパラメタ λ に基づいて、正則化項 $\lambda R(w)$ を計算する（S202）。 $R(w)$ は入力するモデルパラメタから算出され、非特許文献2ではL2

[0020] [数3]

$$R_{L2}(w) = \frac{1}{2} \|w^I\|^2$$

[0021] やGroup Lasso

[0022] [数4]

$$R_{group}(w) = \sum_{j=1}^{N_{I-1}} \|w_j^I\|$$

[0023] が用いられている。正則化項には、損失関数への影響を調整するためのハイパーパラメタ λ を用いる。

[0024] [モデル更新部203]

入力：モデル（更新前）、出力確率分布、正解ユニット番号、正則化項

出力：モデル（更新後）

処理：

モデル更新部203は、正解ユニット番号（教師データにおける正解ラベル）と、正解ユニット番号（教師データにおける正解ラベル）に対応する中間特徴量をニューラルネットワークモデルに入力して得られる出力確率分布

と、正則化項から損失関数を計算し、損失関数 $L(w)=E(w)+\lambda R(w)$ の値を減少させるようにニューラルネットワークモデルを更新する（S 2 0 3）。

先行技術文献

非特許文献

[0025] 非特許文献1：Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patric Nguyen, Tara N. Sainath and Brian Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” IEEE Signal Processing Magazine, Vol. 29, No 6, pp.82-97, 2012.

非特許文献2：T. Ochiai, S. Matsuda, H. Watanabe, and S. Katagiri, “Automatic Node Selection for Deep Neural Networks Using Group Lasso Regularization.” ICASSP, pp. 5485-48, 2017.

発明の概要

発明が解決しようとする課題

[0026] 従来技術 1 では、ニューラルネットワークの学習は指定したパラメタ数をもとにモデルの構築が行われる。すなわち構築されたモデルのサイズは設計者に依存する。しかしながら、このモデルの内部には不要なパラメタが存在しており、ローカルで動作するための音声認識システムを構築するにはモデルサイズや計算量の面でコストがかかるという課題がある（課題 1）。

[0027] この課題 1 に対して従来技術 2 ではモデルを従来通り学習しながら不要なモデルパラメタ（0に近い値）を削除することでモデルサイズを削減する方法が提案されている。通常のL2正則化（上述）ではパラメタ全体の値は小さくなる（行列の0に近い要素が増える）が、行あるいは列ごと削除できないためモデルサイズおよび計算量は削減できない。従来技術 2 ではグループごとのノルムの値を0に近づけるGroup Lasso（上述）を用いることで、行あるいは列をグループとし、学習後にノルムの値が0に近い行あるいは列を削除することでモデルサイズと計算量の削減を実現した。具体的にはGroup Lasso

では行あるいは列をグループとし、図11の(a)に示すようにグループごとにノルムの値を計算したときの頻度を分布とみなし、分布間の境界にあたる値を閾値とし、その閾値よりも小さいノルムとなるグループに該当する行あるいは列に対応するモデルパラメタを削除することでモデルサイズの削減を行う。このとき、Group Lassoではこのノルムの値の頻度の分布を調整できないため、削減するモデルパラメタの数を調整できないという課題がある（課題2）。課題2により、現状のGroup Lassoを用いたニューラルネットワークのモデルサイズ削減方法ではモデルサイズの削減量を調整することは困難である。

[0028] そこで本発明では、モデルサイズの削減量を調整することができるニューラルネットワーク学習装置を提供することを目的とする。

課題を解決するための手段

[0029] 本発明のニューラルネットワーク学習装置は、グループパラメタ生成部と、正則化項計算部と、モデル更新部を含む。

[0030] グループパラメタ生成部は、ニューラルネットワークモデルのモデルパラメタを、任意に定義したグループにグループ分けし、各グループの特徴を表すグループパラメタを生成する。

[0031] 正則化項計算部は、グループパラメタの分布が、分布の特徴を規定するパラメタであるハイパーパラメタによって規定される分布に従うことを仮定して正則化項を計算する。

[0032] モデル更新部は、教師データにおける正解ラベルと、教師データにおける正解ラベルに対応する特徴量をニューラルネットワークモデルに入力して得られる出力確率分布と、正則化項から損失関数を計算し、損失関数の値を減少させるようにニューラルネットワークモデルを更新する。

発明の効果

[0033] 本発明のニューラルネットワーク学習装置によれば、モデルサイズの削減量を調整することができる。

図面の簡単な説明

[0034] [図1]従来技術 1 のニューラルネットワーク学習装置の構成を示すブロック図。

[図2]従来技術 1 のニューラルネットワーク学習装置の動作を示すフローチャート。

[図3]従来技術 2 のニューラルネットワーク学習装置の構成を示すブロック図。

[図4]従来技術 2 のニューラルネットワーク学習装置の動作を示すフローチャート。

[図5]従来技術 2 のスパースモデル更新部の構成を示すブロック図。

[図6]従来技術 2 のスパースモデル更新部の動作を示すフローチャート。

[図7]実施例 1 のニューラルネットワーク学習装置の構成を示すブロック図。

[図8]実施例 1 のニューラルネットワーク学習装置の動作を示すフローチャート。

[図9]実施例 1 のスパースモデル更新部の構成を示すブロック図。

[図10]実施例 1 のスパースモデル更新部の動作を示すフローチャート。

[図11]従来技術 2 と実施例 1 のニューラルネットワーク学習装置の違いを説明する概念図であって、図 1 1 (a) は、従来技術 2 のニューラルネットワーク学習装置のモデルパラメタ削除の概要を示す図、図 1 1 (b) は、実施例 1 のニューラルネットワーク学習装置のモデルパラメタ削除の概要を示す図。

発明を実施するための形態

[0035] 以下、本発明の実施の形態について、詳細に説明する。なお、同じ機能を有する構成部には同じ番号を付し、重複説明を省略する。

実施例 1

[0036] 実施例 1 のニューラルネットワーク学習装置では、従来技術 2 における正則化技術を改善した。従来技術 2 ではモデルサイズの削減量を定義できなかったが、実施例 1 のニューラルネットワーク学習装置ではモデルサイズの削減量に影響するパラメタを導入することにより、モデルサイズの削減量を調

整することができる。

[0037] 以下、図7を参照して実施例1のニューラルネットワーク学習装置300の構成について説明する。同図に示すように、ニューラルネットワーク学習装置300は、中間特徴量抽出部101と、出力確率分布計算部102と、スパースモデル更新部301を含み、中間特徴量抽出部101と、出力確率分布計算部102については、従来技術1、従来技術2の同名の構成要件と同じ動作を実行する。以下、図8を参照して、スパースモデル更新部301の動作について説明する。

[0038] [スパースモデル更新部301]

入力：モデル（更新前）、出力確率分布、正解ユニット番号

出力：スパースなモデル（更新後）

処理：

スパースモデル更新部301は、グループパラメタと正則化項 $\lambda R(w)$ を生成し、正解ユニット番号と出力確率分布とグループパラメタと正則化項 $\lambda R(w)$ から損失関数を計算し、損失関数の値を減少させるようにモデルを更新し、スパースなモデルを出力する（S301）。

[0039] 以下、図9を参照して、スパースモデル更新部301の詳細について説明する。同図に示すように、スパースモデル更新部301は、グループパラメタ生成部302と、正則化項計算部303と、モデル更新部203を含む構成である。モデル更新部203は、従来技術2における同名の構成要件と同じ動作を実行する。以下、図10を参照して、スパースモデル更新部301内の各構成要件の動作について説明する。

[0040] [グループパラメタ生成部302]

入力：モデル（更新前）、グループの定義（行、列といった具体的なグルーピングの方法）

出力：グループパラメタ

処理：

グループパラメタ生成部302は、入力したモデル（更新前）のモデルパ

ラメタを、上記グループの定義（行、列といった具体的なグルーピングの方法）によって任意に定義したグループにグループ分けし、各グループの特徴を表すグループパラメタを生成する（S 3 0 2）。別の表現では、グループパラメタ生成部 3 0 2 は、入力したモデルパラメタに対して、上記グループの定義によってグループを定義し、グループ空間における分布に基づくグループパラメタを取得する。グループパラメタの具体例としては、例えば、モデルパラメタを行列とし、グループの定義により、グループをモデルパラメタの行列のうちの行または列と定義したときに、行ベクトルまたは列ベクトルごとのノルムの値などである。

[0041] [正則化項計算部 3 0 3]

入力：グループパラメタ、ハイパーパラメタ

出力：正則化項

処理：

正則化項計算部 3 0 3 は、グループパラメタの分布が、分布の特徴を規定するパラメタであるハイパーパラメタによって規定される分布に従うことを仮定して正則化項を計算する（S 3 0 3）。正則化項計算部 3 0 3 は、以下に示すようなグループパラメタ w_g が分布に従うことを仮定した正則化項 $R_{proposed}(w)$ を用いる点において、正則化項計算部 2 0 2 と異なる処理を実行する。

[0042] [数5]

$$R_{proposed}(w) = \sum_{g \in G} \log \left\{ \sum_{j=1}^m \alpha_j \exp \left(- \frac{(\|w_g\|_2 - \mu_j)^2}{2\sigma_j^2} \right) \right\}$$

[0043] w_g は任意に決めることができるグループ（ベクトルあるいは行列）における任意のパラメタ（例えばノルム）であり、従来技術 2 に合わせるとモデルパラメタの行列の行ベクトルあるいは列ベクトルにおける任意のパラメタ（例えばノルム）を示す。上式の括弧 $\{*\}$ 内は混合ガウス分布を表しており、 j および m はグループパラメタの分布を仮定した場合の分布の混合数を表す。混合重み α_j 、平均 μ_j 、分散 σ_j はグループパラメタの分布を調整するためのハイパ

一パラメタであり、これらのハイパーパラメタを調整（例：混合重み α_j の比を変えることで平均 μ_j に属するパラメタの重要度を調整）することでモデルサイズの削減量を調整することが可能となる。また、上式では混合ガウス分布を仮定したが実際はガウス分布以外に任意の分布を組み合わせることが可能である。ラプラス分布とガウス分布を組み合わせた場合の正則化項を以下に示す。

[0044] [数6]

$$R'_{proposed}(\mathbf{w}) = \sum_{g \in G} \log \left\{ \sum_{j=1}^m \alpha_j \exp \left(-\frac{(\|\mathbf{w}_g\|_2 - \mu_j)^2}{2\sigma_j^2} \right) + \sum_{k=1}^n \beta_k \exp \left(-\frac{|\|\mathbf{w}_g\|_2 - \mu'_k|}{\sigma'_k} \right) \right\}$$

[0045] 第二項は混合ラプラス分布であり、混合重み β_k 、平均 μ'_k 、分散 σ'_k も第一項のハイパーパラメタ同様にモデルサイズの削減量を調整するハイパーパラメタである。上式では微分不可能な点を持つラプラス分布を用いることも可能であることを示しているが実利用では全区間で微分可能であることが望ましい。最後に任意の分布関数 $F(*)$ を用いて一般化した場合を以下に示す。

[0046] [数7]

$$R''_{proposed}(\mathbf{w}) = \sum_{g \in G} \log \left\{ \sum_{j=1}^m \alpha_j F \left(\frac{|\|\mathbf{w}_g\|_2 - \mu_j|}{\sigma_j} \right) \right\}$$

[0047] 混合重み α_j 、平均 μ_j 、分散 σ_j は任意の分布関数のハイパーパラメタを表す。

[0048] 従来技術2におけるGroup Lassoによるグループ正則化は行や列をグループとしているがグループパラメタの分布は仮定しておらず、分布の形状などの調整はできない。従来技術2におけるGroup Lassoによるグループ正則化では、図11(a)に示すように、モデルの更新に伴って分布の形状は一通りに収束していくため、削除対象となるグループと残すグループの区分の方法が1通りに限定される。そのため、Group Lassoはグループ空間上におけるモデルパラメタの正則化は可能であるが、グループパラメタの分布を調整するような仕組みが定義されていないため分布の大きさなどを調整できず、不要なモデルパラメタの量を調整することが不可能である。

[0049] 一方、本実施例のニューラルネットワーク学習装置300によるグループ正則化では、図11(b)に示すように、グループパラメタの分布の特徴を規定するハイパーパラメタ（例：混合重み、平均、分散など）を生成したため、削除対象となるグループ、残すグループの各分布の形状をカスタマイズすることができ、削除対象となるモデルパラメタの量を調整することができる。

[0050] <効果>

本実施例のニューラルネットワーク学習装置300により作成したニューラルネットワークを用いることでモデルサイズの削減量を調整することができ、認識精度を維持しながらGroup Lassoよりもモデルのサイズを削減するといったカスタマイズが可能となるため、ニューラルネットワークを用いたモデルをローカルのシステムに組み込む上でモデルサイズおよび計算量の面で非常に効果的である。

[0051] <補記>

本発明の装置は、例えば単一のハードウェアエンティティとして、キーボードなどが接続可能な入力部、液晶ディスプレイなどが接続可能な出力部、ハードウェアエンティティの外部に通信可能な通信装置（例えば通信ケーブル）が接続可能な通信部、CPU（Central Processing Unit、キャッシュメモリやレジスタなどを備えていてもよい）、メモリであるRAMやROM、ハードディスクである外部記憶装置並びにこれらの入力部、出力部、通信部、CPU、RAM、ROM、外部記憶装置の間のデータのやり取りが可能なように接続するバスを有している。また必要に応じて、ハードウェアエンティティに、CD-ROMなどの記録媒体を読み書きできる装置（ドライブ）などを設けることとしてもよい。このようなハードウェア資源を備えた物理的実体としては、汎用コンピュータなどがある。

[0052] ハードウェアエンティティの外部記憶装置には、上述の機能を実現するために必要となるプログラムおよびこのプログラムの処理において必要となるデータなどが記憶されている（外部記憶装置に限らず、例えばプログラムを

読み出し専用記憶装置であるROMに記憶しておくこととしてもよい)。

また、これらのプログラムの処理によって得られるデータなどは、RAMや外部記憶装置などに適宜に記憶される。

[0053] ハードウェアエンティティでは、外部記憶装置（あるいはROMなど）に記憶された各プログラムとこの各プログラムの処理に必要なデータが必要に応じてメモリに読み込まれて、適宜にCPUで解釈実行・処理される。その結果、CPUが所定の機能（上記、…部、…手段などと表した各構成要件）を実現する。

[0054] 本発明は上述の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で適宜変更が可能である。また、上記実施形態において説明した処理は、記載の順に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されるとしてもよい。

[0055] 既述のように、上記実施形態において説明したハードウェアエンティティ（本発明の装置）における処理機能をコンピュータによって実現する場合、ハードウェアエンティティが有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記ハードウェアエンティティにおける処理機能がコンピュータ上で実現される。

[0056] この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等のようなものでもよい。具体的には、例えば、磁気記録装置として、ハードディスク装置、フレキシブルディスク、磁気テープ等を、光ディスクとして、DVD (Digital Versatile Disc)、DVD-RAM (Random Access Memory)、CD-ROM (Compact Disc Read Only Memory)、CD-R (Recordable) / RW (ReWritable) 等を、光磁気記録媒体として、MO (Magneto-Optical disc) 等を、半導体メモリとしてEEPROM (Ele

ctronically Erasable and Programmable-Read Only Memory) 等を用いることができる。

[0057] また、このプログラムの流通は、例えば、そのプログラムを記録したDVD、CD-ROM等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

[0058] このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、一旦、自己の記憶装置に格納する。そして、処理の実行時、このコンピュータは、自己の記録媒体に格納されたプログラムを読み取り、読み取ったプログラムに従った処理を実行する。また、このプログラムの別の実行形態として、コンピュータが可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるASP (Application Service Provider) 型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

[0059] また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、ハードウェアエンティティを構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

請求の範囲

- [請求項1] ニューラルネットワークモデルのモデルパラメタを、任意に定義したグループにグループ分けし、各グループの特徴を表すグループパラメタを生成するグループパラメタ生成部と、
- 前記グループパラメタの分布が、分布の特徴を規定するパラメタであるハイパーパラメタによって規定される分布に従うことを仮定して正則化項を計算する正則化項計算部と、
- 教師データにおける正解ラベルと、前記教師データにおける前記正解ラベルに対応する特徴量を前記ニューラルネットワークモデルに入力して得られる出力確率分布と、前記正則化項から損失関数を計算し、損失関数の値を減少させるように前記ニューラルネットワークモデルを更新するモデル更新部を含む
- ニューラルネットワーク学習装置。
- [請求項2] 請求項1に記載のニューラルネットワーク学習装置であって、
- 前記モデルパラメタを行列とし、前記グループを前記モデルパラメタの行列のうちの行または列とし、前記グループパラメタを行ベクトルまたは列ベクトルのノルムとした
- ニューラルネットワーク学習装置。
- [請求項3] 請求項1または2に記載のニューラルネットワーク学習装置であって、
- 前記ハイパーパラメタは、混合重み、平均、分散のうち少なくとも何れか一つを含む
- ニューラルネットワーク学習装置。
- [請求項4] ニューラルネットワーク学習装置が実行するニューラルネットワーク学習方法であって、
- ニューラルネットワークモデルのモデルパラメタを、任意に定義したグループにグループ分けし、各グループの特徴を表すグループパラメタを生成するステップと、

前記グループパラメタの分布が、分布の特徴を規定するパラメタであるハイパーパラメタによって規定される分布に従うことを仮定して正則化項を計算するステップと、

教師データにおける正解ラベルと、前記教師データにおける前記正解ラベルに対応する特徴量を前記ニューラルネットワークモデルに入力して得られる出力確率分布と、前記正則化項から損失関数を計算し、損失関数の値を減少させるように前記ニューラルネットワークモデルを更新するステップを含む

ニューラルネットワーク学習方法。

[請求項5]

請求項4に記載のニューラルネットワーク学習方法であって、

前記モデルパラメタを行列とし、前記グループを前記モデルパラメタの行列のうちの行または列とし、前記グループパラメタを行ベクトルまたは列ベクトルのノルムとした

ニューラルネットワーク学習方法。

[請求項6]

請求項4または5に記載のニューラルネットワーク学習方法であって、

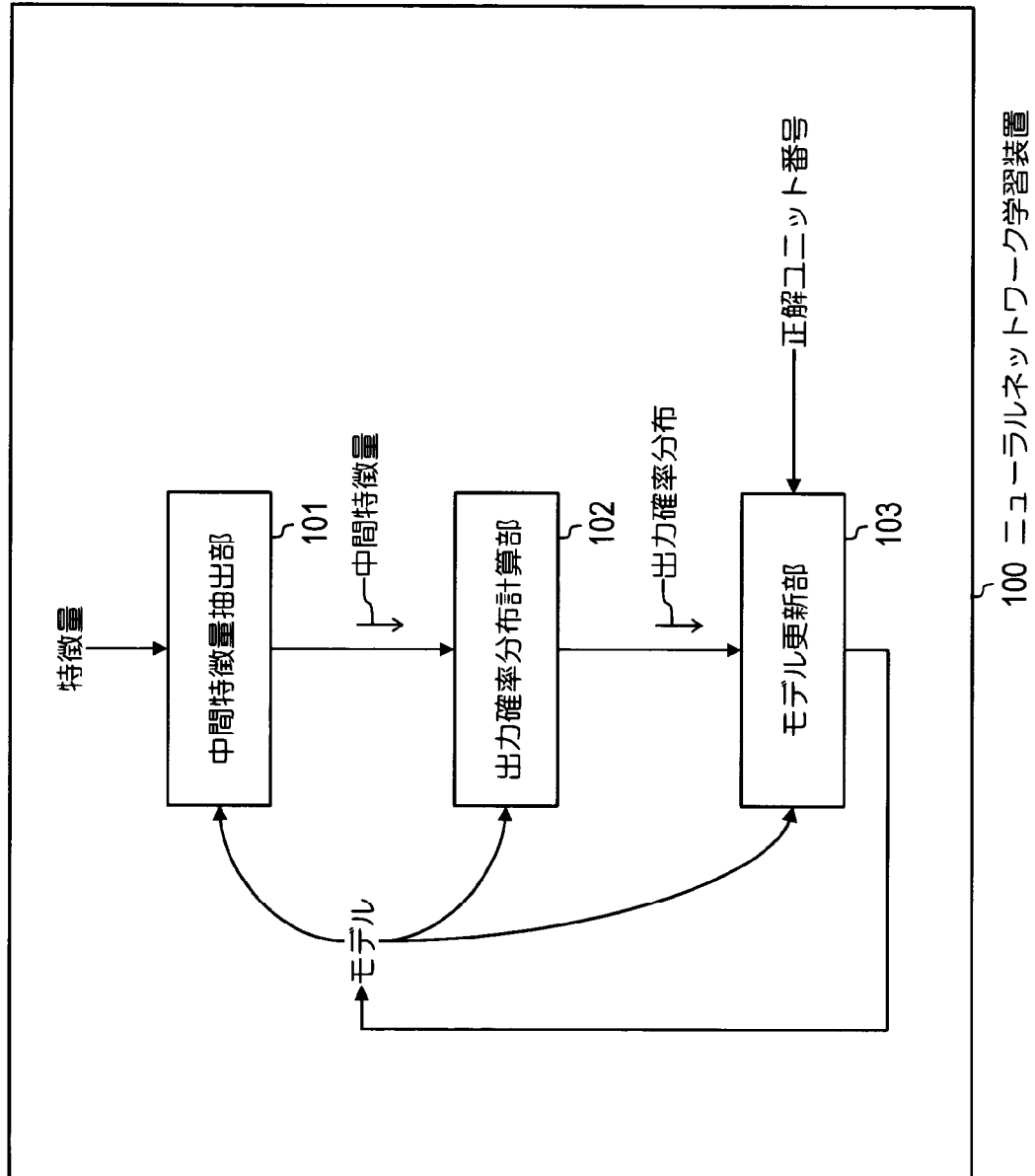
前記ハイパーパラメタは、混合重み、平均、分散のうち少なくとも何れか一つを含む

ニューラルネットワーク学習方法。

[請求項7]

コンピュータを、請求項1から3の何れかに記載のニューラルネットワーク学習装置として機能させるプログラム。

[図1]



[図2]

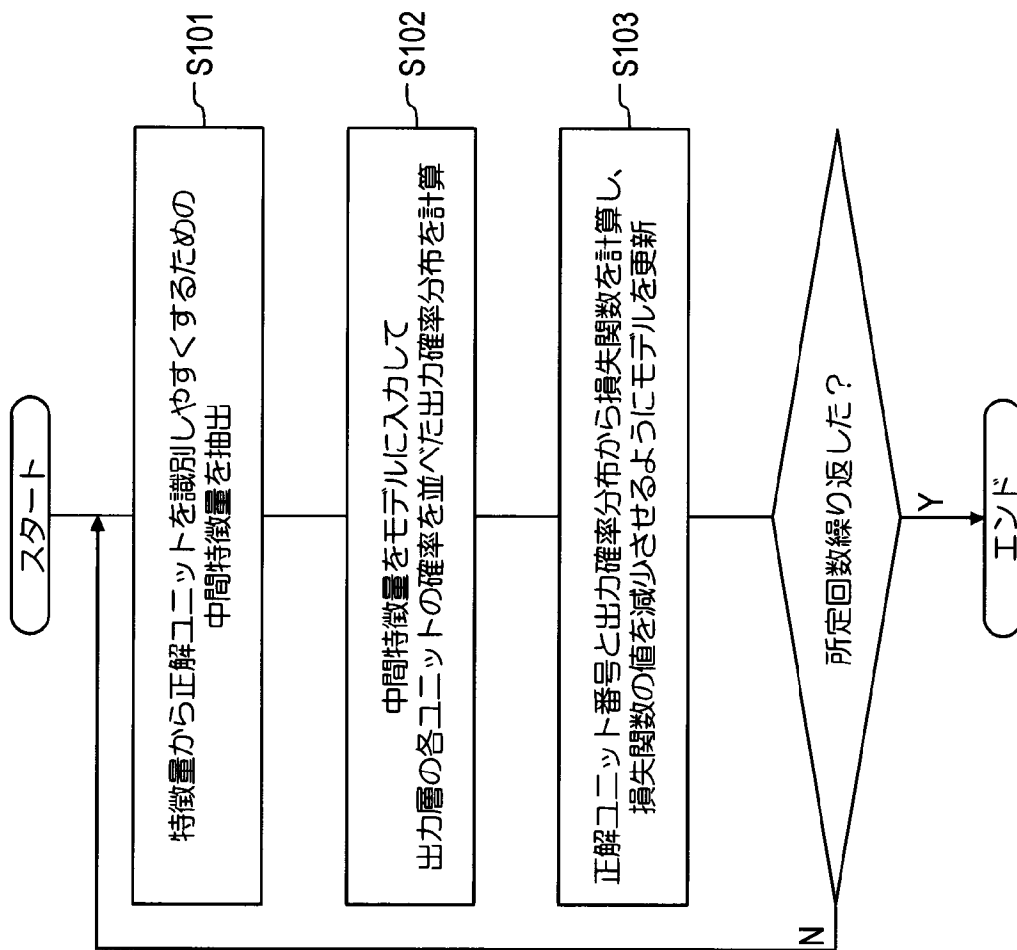
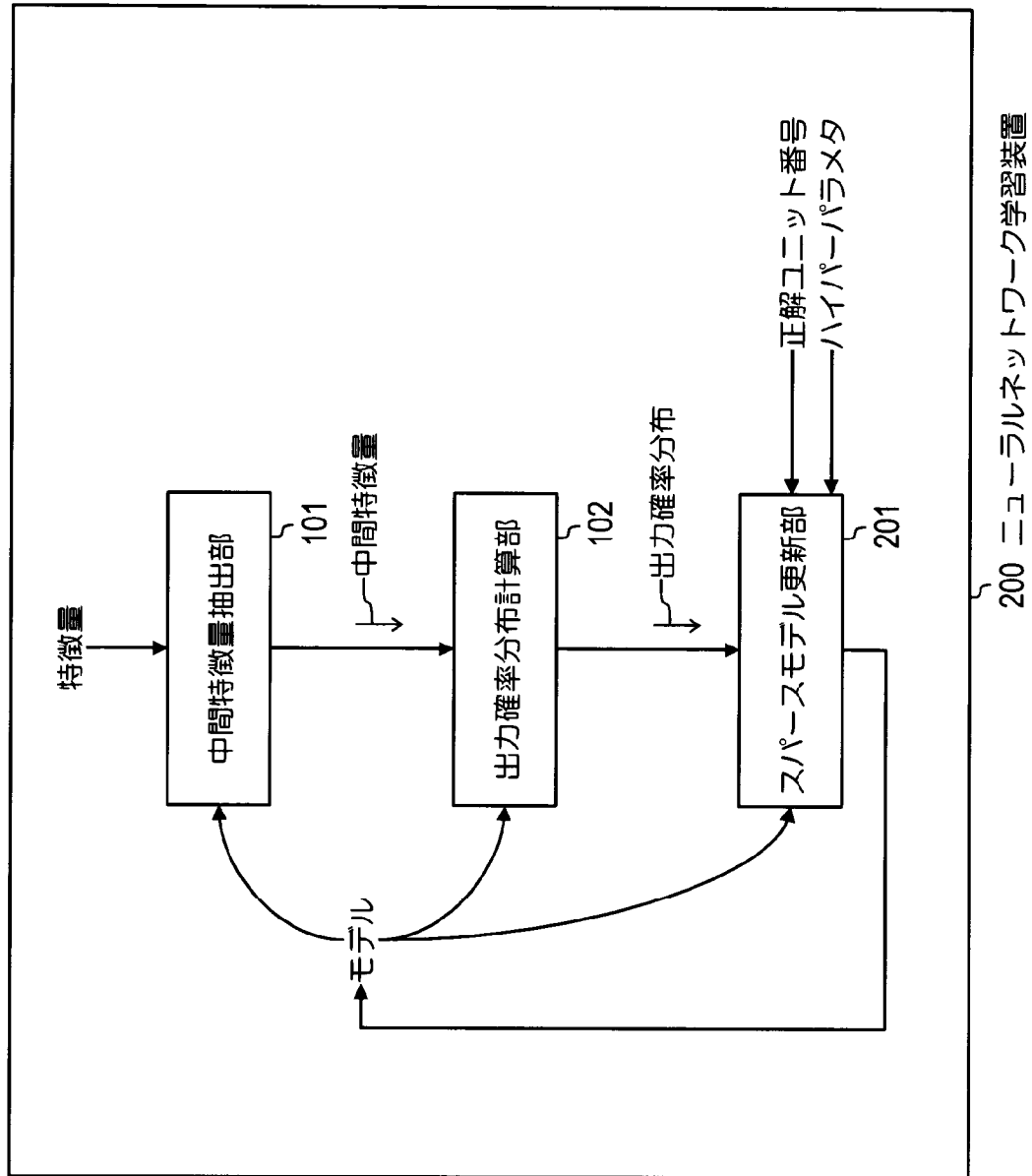
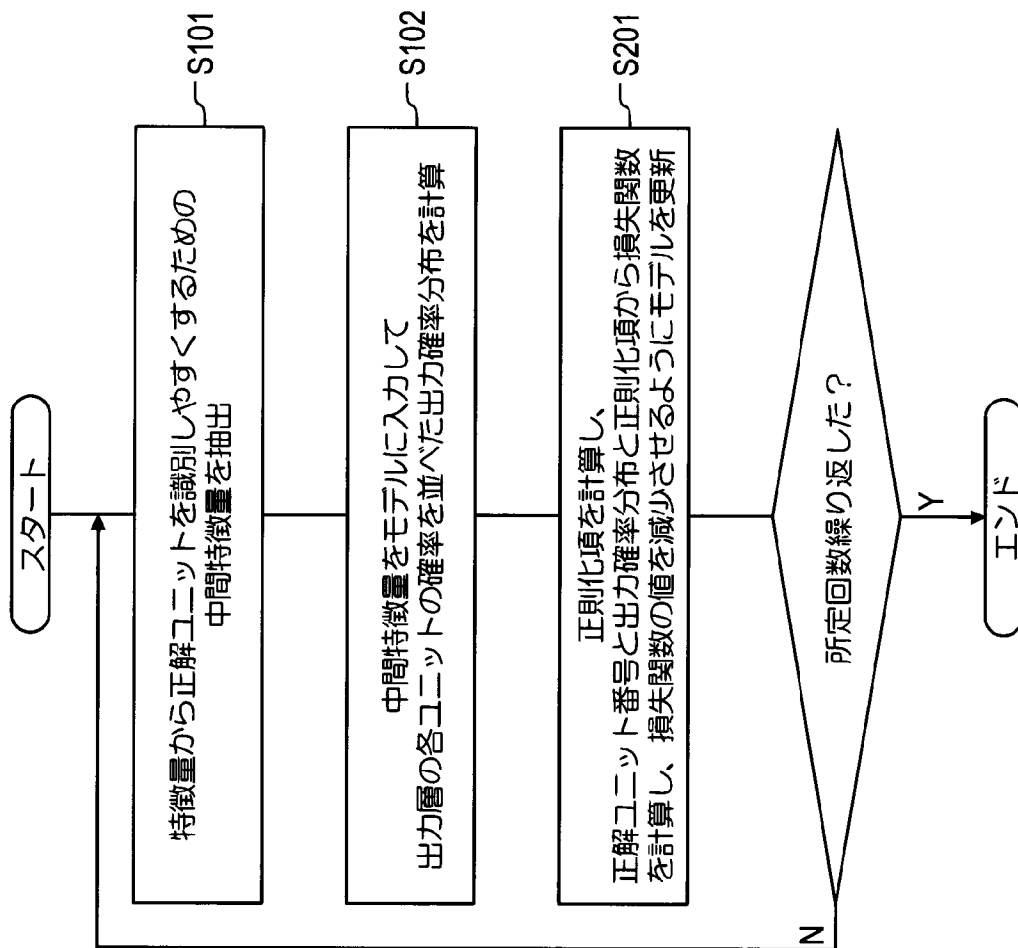


図2

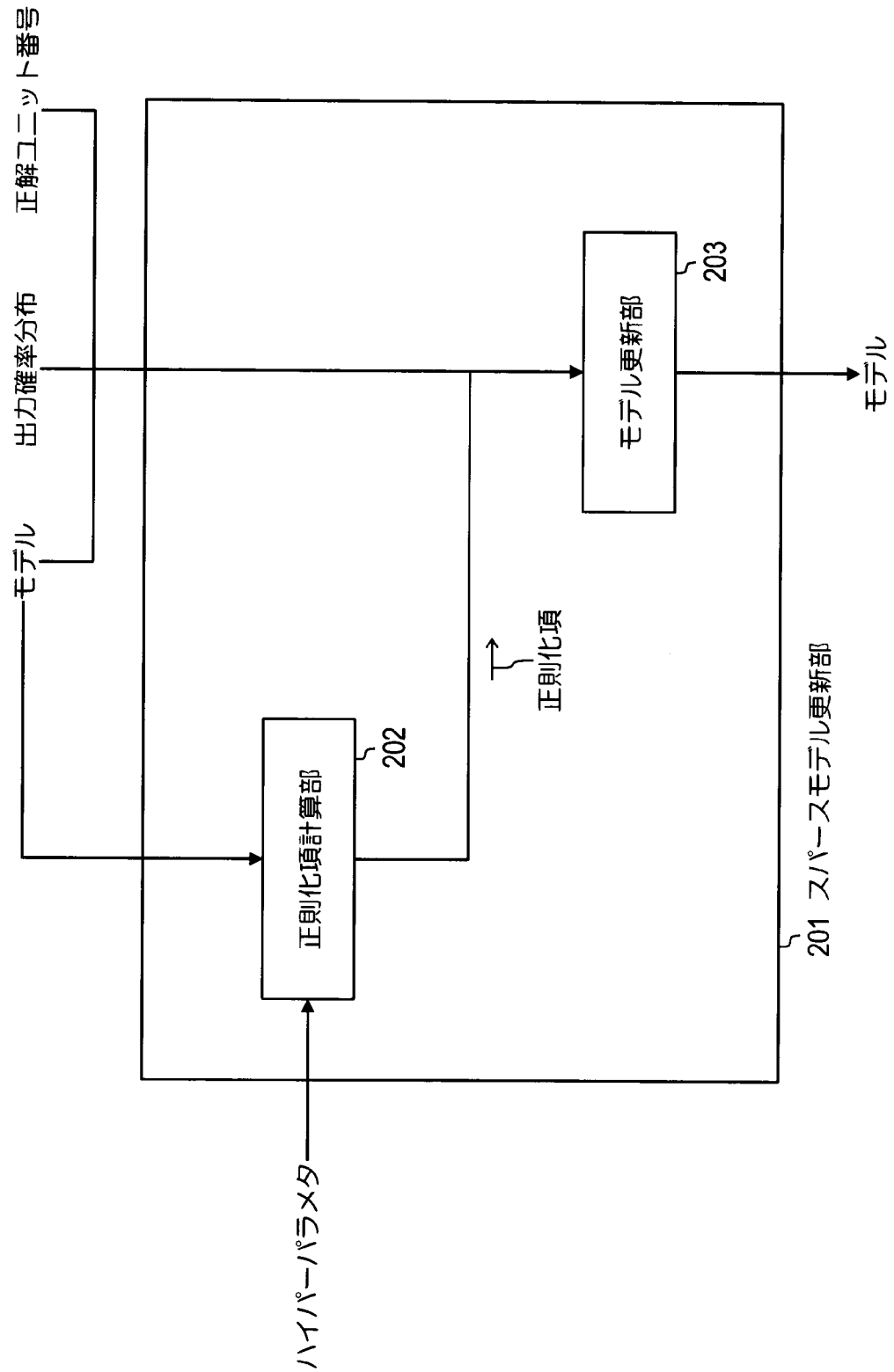
[図3]



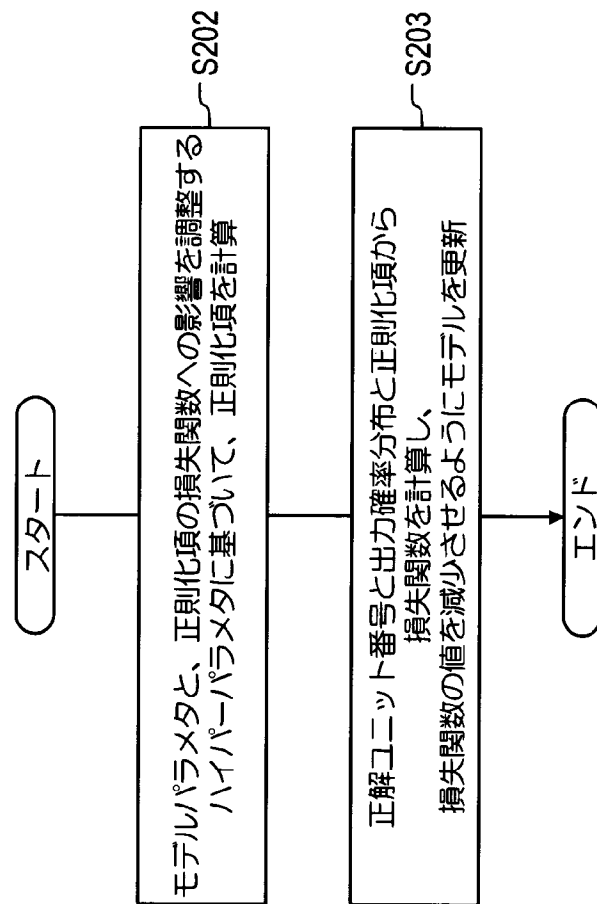
[図4]



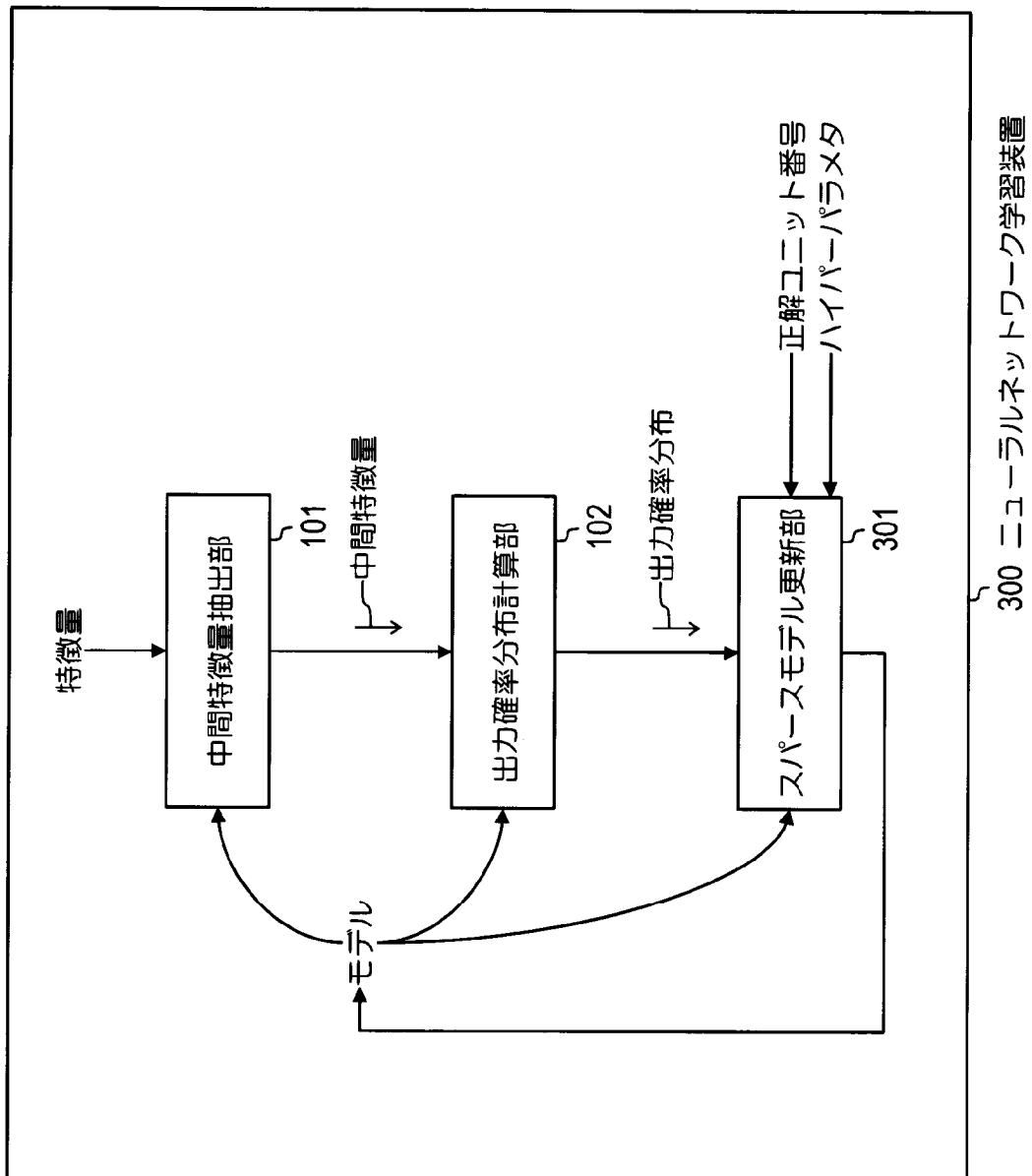
[図5]



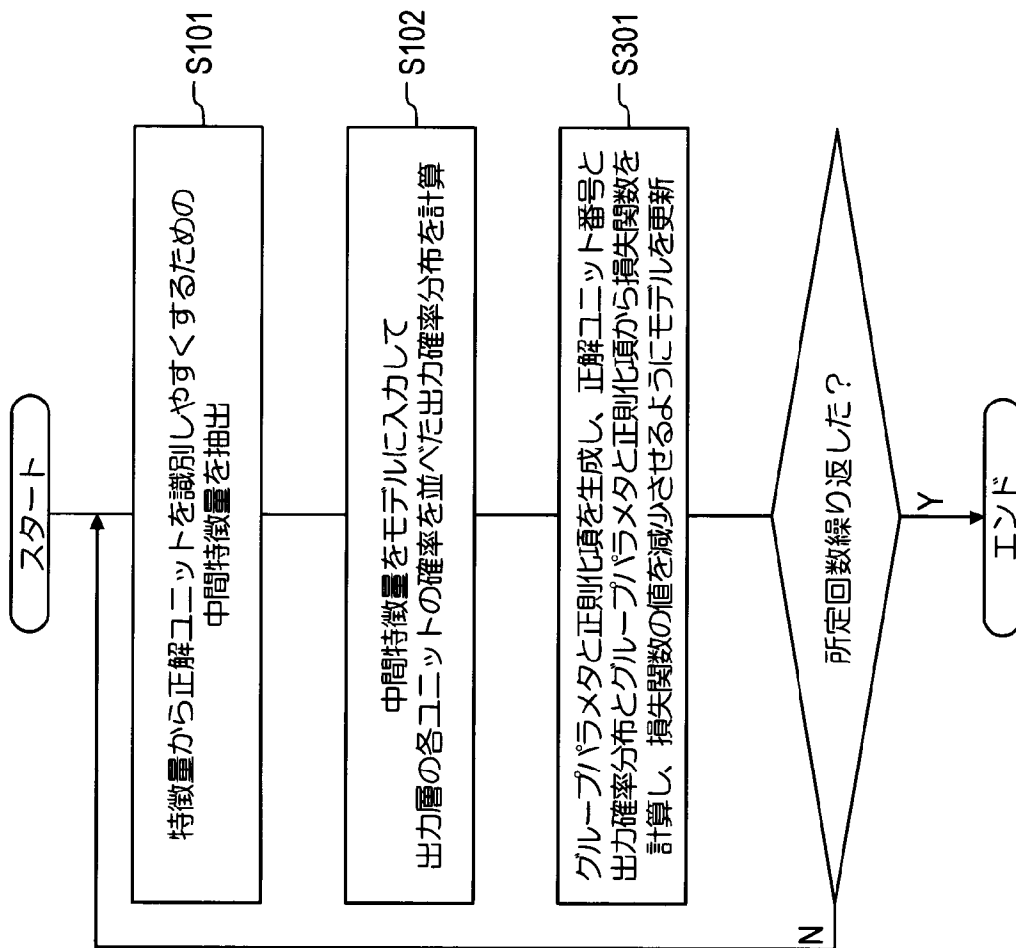
[図6]



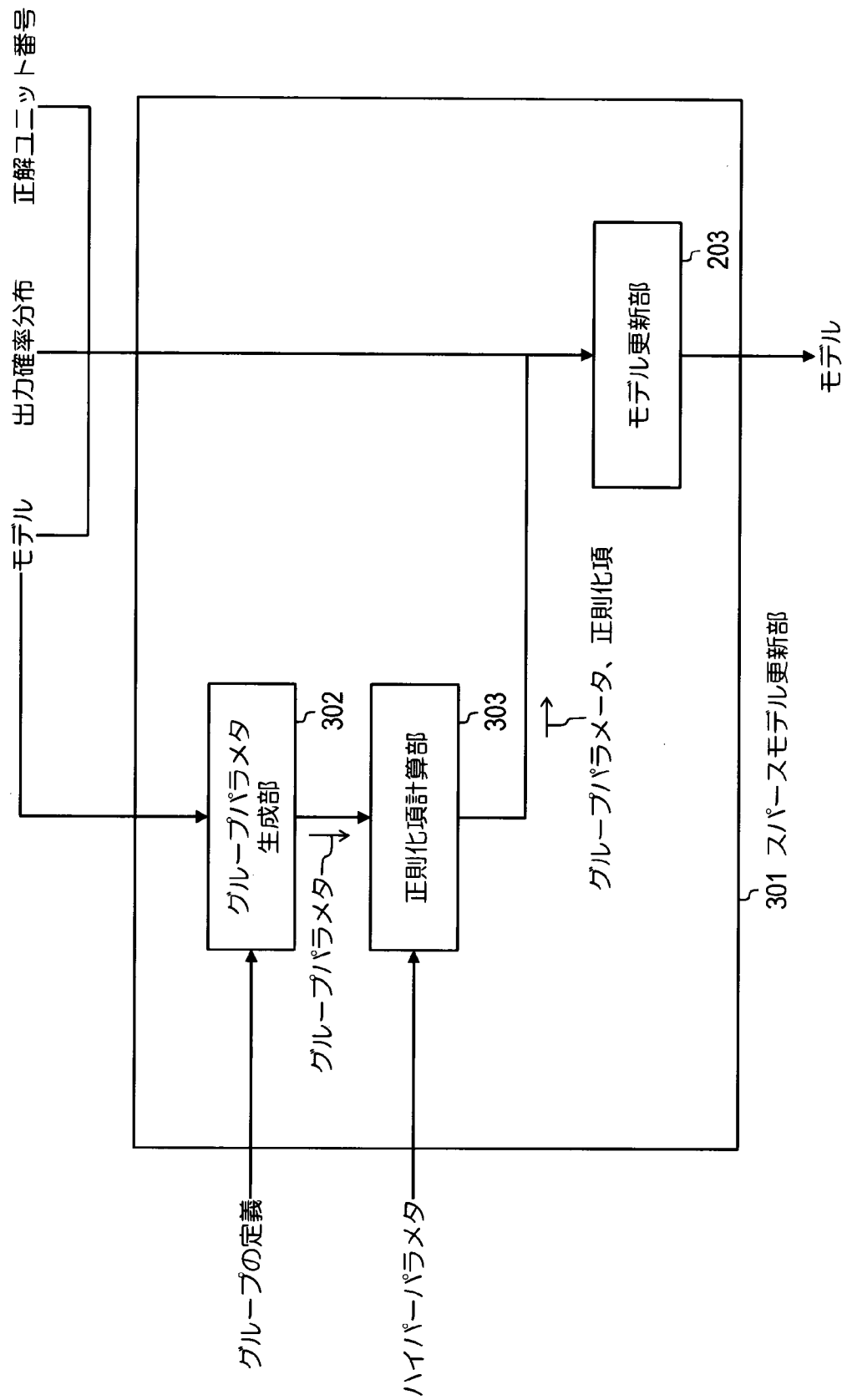
[図7]



[図8]



[図9]



[図10]

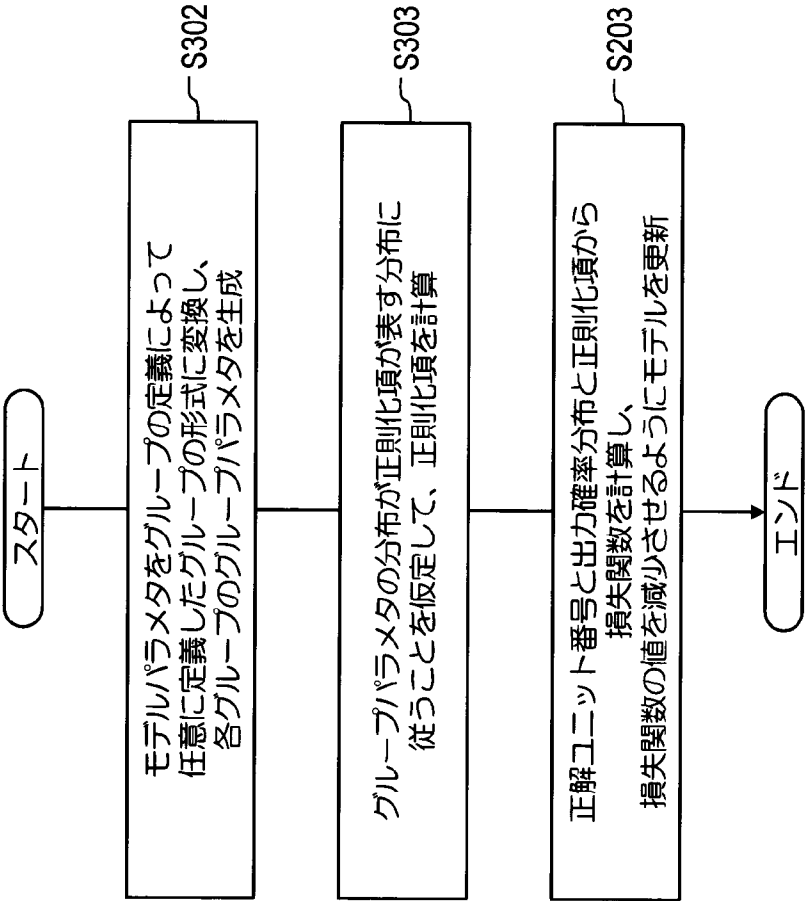


図10

[図11]

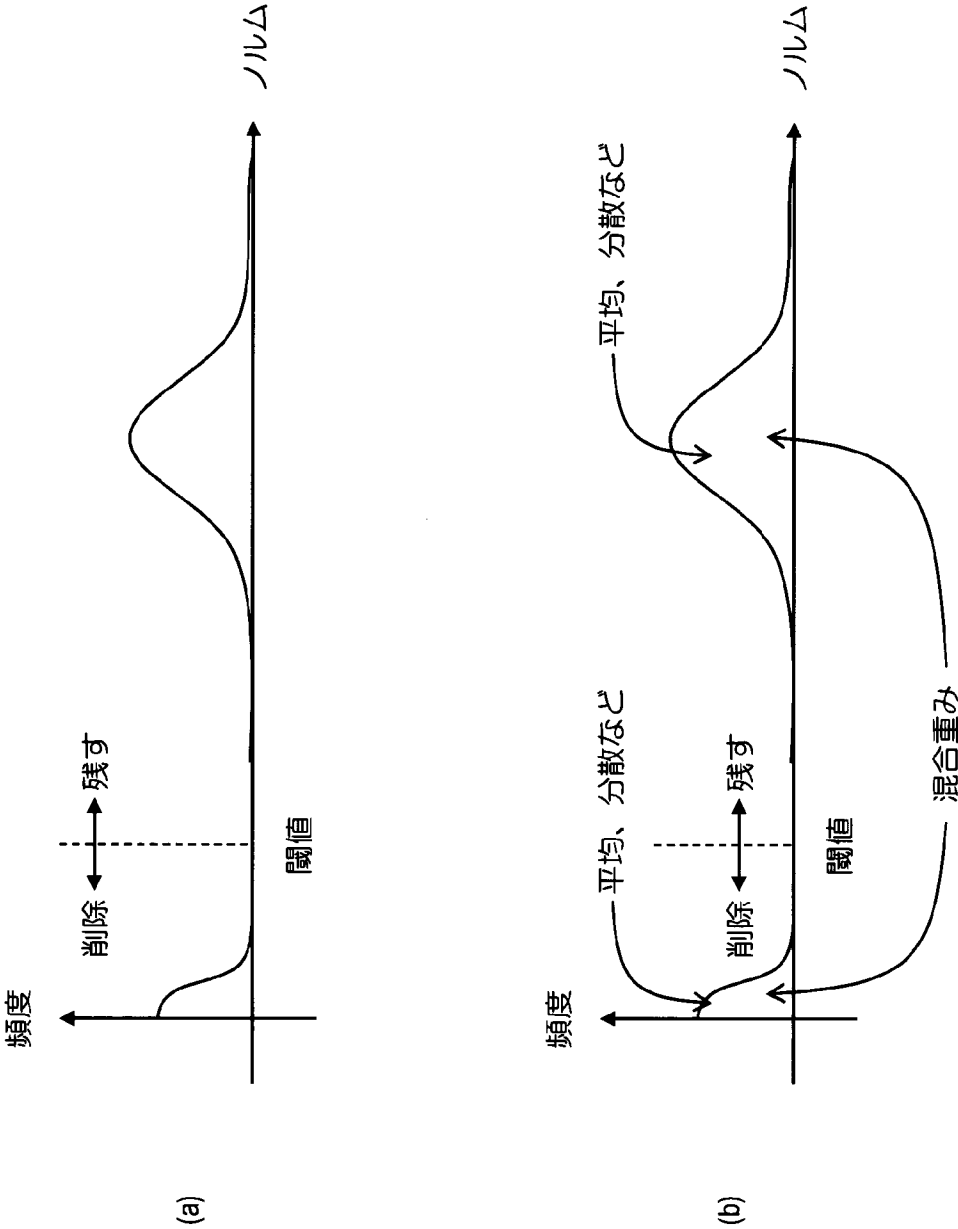


図11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/017216

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06N3/08 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int.Cl. G06N3/08

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2019

Registered utility model specifications of Japan 1996-2019

Published registered utility model applications of Japan 1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCARDAPANE, S. et al., Group sparse regularization for deep neural networks, v1, 02 July 2016, pp. 1-10, [online], [retrieved on 10 June 2019], Internet: <URL:https://arxiv.org/pdf/1607.00485.pdf>, <DOI:10.1016/j.neucom.2017.02.029>	1-7

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 10.06.2019	Date of mailing of the international search report 18.06.2019
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/017216

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	XU, X. F. et al., Bayesian variable selection and estimation for group lasso, v1, 03 December 2015, pp. 909-936, [online], [retrieved on 10 June 2019], Internet: <URL:https://arxiv.org/pdf/1512.01013.pdf>, <DOI:10.1214/14-BA929>	1-7

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06N3/08(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl. G06N3/08

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	SCARDAPANE, Simone et al., Group Sparse Regularization for Deep Neural Networks, v1, 2016.07.02, p.1-10, [オンライン]、[検索日：2019.06.10]、インターネット： <URL:https://arxiv.org/pdf/1607.00485.pdf>、 <DOI:10.1016/j.neucom.2017.02.029>	1-7

☒ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

10.06.2019

国際調査報告の発送日

18.06.2019

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

福西 章人

5 B

4687

電話番号 03-3581-1101 内線 3545

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	XU, Xiaofan et al., Bayesian Variable Selection and Estimation for Group Lasso, v1, 2015.12.03, p.909-936, [オンライン]、[検索日：2019.06.10]、インターネット： <URL:https://arxiv.org/pdf/1512.01013.pdf>、 <DOI:10.1214/14-BA929>	1-7