2C3-OS-06b-1

Deep Learningが獲得する特徴表現の理解と利用に向けた中間層情報の活用

Exploiting the Hidden Layer Information Toward the Understanding and Utilization of Feature Representations Obtained from Deep Learning

菊田　遥平 *1　野村　眞平 *2　吉永　恵一 *2　小林　秀 *1　神津　友武 *1
Yohei Kikuta　Shinpei Nomura　Keiichi Yoshinaga　Shu Kobayashi　Tomotake Kozu

*1有限責任監査法人トーマツ　デロイトアナリティクス **
Deloitte Analytics, Deloitte Touche Tohmatsu LLC.

*2株式会社リクルート住まいカンパニー
Recruit Sumai Company Ltd.

Deep Learning has recently been a hot topic due to the high performance on tasks such as image recognition and speech recognition. However, it seems that the understanding and utilization of hidden layer representations are limited and insufficient. In this paper we propose an approach that uses hidden layer information, especially learned weights, to quantify the importance and similarity of hidden layer nodes. We then investigate the set of input nodes having large contributions to the output and the optimization of numbers of hidden layers. The proposed approach is applied to the data of SUUMO, a housing information site in Japan. We show the results that how hidden layers acquire the features of input nodes and how the features and the discriminate accuracy change subject to the number of hidden layers.

1. はじめに

近年, 大量のデータを学習することで高い識別精度を発揮する Deep Learning が注目を集めている. 各種コンペティションや画像もしくは音声認識を対象としたビジネス上の応用などが顕著な成功例であるが, その適用範囲や可能性は今後も継続するであろうデータの質の向上と量の増加や計算機の発展に伴いさらに広がっていくものと期待される. Deep Learning が注目されている理由の一つはその識別精度の高さにあるが, 従来までの機械学習と本質的に異なる点は, Google の猫 [Le 12] に代表されるような自動的な特徴表現の獲得にある. この特徴表現とは Deep Learning のネットワーク構造がデータを学習した結果得られる中間層情報そのものであり, 中間層情報を活用することが Deep Learning が獲得する特徴表現の理解と利用へと繋がる.

中間層情報が活用できるようになれば, そこから得られる段階的な特徴量を用いた応用を推し進めることができる. 例えば, 各層での特徴量を基にデータクラスタリングを行うことでクラスタリングの粒度調整が実施できる. これにより, 表記揺れ→同義語→関連語とまとめるような逐次的な辞書構築やアクセスログの情報だけを用いたユーザ層の段階的なセグメント化など, 様々なサービスが展開可能となる. また, 中間層情報により分析結果の理由づけや解釈が進めば, 単に精度だけでなく説明性が求められるビジネス利用において有用である.

しかしながら, 中間層情報の把握と活用はまだ限定的であり, 一部の非構造化データを対象にした議論がなされているに過ぎない. 原因として, 中間層情報を扱うためにライブラリやツールに一定の習熟が必要なこと, 中間層情報はそのままでは意味の解釈が難しいこと, モデルパラメタである中間層の層数やノード数はアプリオリに決めるしかなく指針がないこと, などが挙げられる.

連絡先: 菊田遥平, 有限責任監査法人トーマツ　デロイトアナリティクス, yohei.kikuta@tohmatsu.co.jp

** 本研究の内容は有限責任監査法人トーマツの公式見解を示すものではありません.

本研究では中間層情報の活用に向けた一つの試みとして, 学習後に得られる重み行列から中間層ノードの重要度と類似度を定め, それを基にした解析を行う. 重要度は重み行列成分の値の大きさで定義し, 重要度から出力ノードへの寄与が大きい中間層ノードを特定し, その中間層ノードにどのような説明変数が強く結びつくかを調べる. 類似度は重み行列の各行をベクトルとして扱った際のベクトル間距離の逆数で定義し, 類似度を用いて中間層ノード数を最適化するクラスタリングを実施し, そのノード数によって中間層ノードに結びつく説明変数や判別精度にどのような変化が生じるのかを調べる. 精度を第一義的な対象とするのではなく, 中間層情報の活用に関する知見を得ることが本研究の目的である.

本研究の提案手法を不動産ポータルサイトのデータに適用した結果, 出力ノードへの影響が大きい中間層ノードが不動産物件閲覧の際の典型的なエリアを特徴として抽出することが明らかになった. また, 中間層ノード数を最適化して少なくしていくにつれて, 抽出される特徴は説明変数が複雑に関係し合って目的変数とは直接の関係が薄く乖離があるものになっていく様子が観測された. 学習したモデルの識別精度に関しては, 中間層ノード数の最適化により一定の向上が見られた.

2. 解析手法

本研究では, 二値判別問題を Deep Learning で解き, 学習の結果として得られる中間層情報を取り出して以下のステップで解析を実行する.

・重要度に基づいて寄与の大きい中間層ノードを抽出し, どのような説明変数がそれらに強く結びつくかを調べる.
・類似度に基づいて中間層のノード数を最適化する.
・上記過程を繰り返してその変化を追う.

なお, 判別精度に関しては出力ノードの値を用いたゲインチャートで検証を行う. ここでは, 上の二つのステップに関して具体的な手法を以下で説明する.

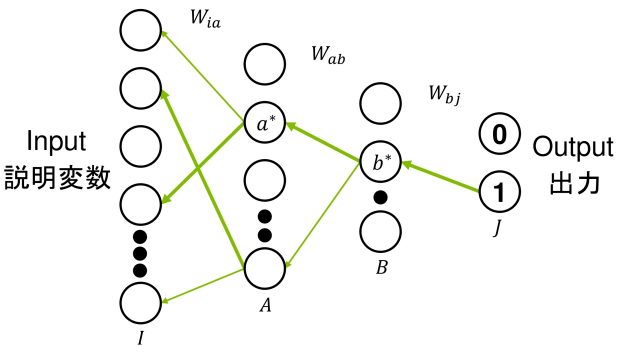

図 1: 二値判別モデルの出力ノードから重み行列を基に寄与の大きいノードを辿っていく概念図. W_{mn} は M 層と N 層をつなぐ重み行列で, 矢印の太さは重みの大きさを表す.

2.1 重要ノード特定と説明変数のグルーピング観測

出力ノードへの寄与が大きい重要ノードを特定するために, 直接的な方法として図 1 のように出力層から重み行列を辿って寄与の大きいノードを明らかにしていく方法を採用する. A 層→ B 層とつなぐ重み行列を W_{ab}(A 行 B 列) と書くとき, 出力層に近い B 層のある特定のノード ($b = b^*$) に対する A 層の特定のノード ($a = a^*$) の重要度を $W_{a^*b^*}$ と定義する.

説明変数まで遡ることで重要度の高い中間層ノードに強く結びつく説明変数の組が見て取れるので, 中間層がデータのどのような特徴を抽出しネットワーク構造を構築するのかを観察することができる.

2.2 中間層ノード数の最適化

現状においては中間層ノード数はアプリオリに決定するしかない. 多くの場合, 精度や学習時間との兼ね合いで発見的に決定されるものであり, 標準的な方法は確立されていない. 本研究では精度という観点ではなく, ノード同士の類似度に基づき似ているノードが存在する場合は余分なノードがあると解釈してノード数を減らすという方法でノード数を最適化する. 類似度を定義するには重み行列が属する空間に距離を導入する必要があるが, ここでは Euclid 計量を入れてベクトル間距離の逆数で類似度を定義することにする.

重み行列 W_{ab} の転置行列 W_{ba}^T の各行を $\vec{w}_b$ と記し B 層のノードを原点としたときの A 層のノードが位置する座標と解釈する. これにより B 層のノードから見たとき同じように結びついている A 層のノードは近い座標に存在していることになる. この座標を用いて kmeans 法によりクラスタリングを実施して似ているノードをまとめ, Gap 統計量 [Tibshirani 01] を導入して最適クラスタリング数を決定する. Gap 統計量の定義は以下の通り.

$$Gap(k) = E_{X^*}[\log D_k(X^*)] - \log D_k(X). \tag{1}$$

ここで, $X = \{\vec{w}_b | b = 1, \ldots, B\}$, X^* はランダムに生成した同じサイズのデータセット, E[] は平均である. $D_k(X)$ は以下で定義される.

$$D_k(X) = \sum_{r=1}^{k} \sum_{i,j}^{|X_r|} \frac{1}{2|X_r|} g_{\alpha\beta}(x_{ri}^{\alpha} - x_{rj}^{\alpha})(x_{ri}^{\beta} - x_{rj}^{\beta}). \tag{2}$$

ここで, データセットは kmeans 法により k 個に分割されていて ($X = \bigcup_{r=1}^{k} X_r$ where $X_m \cap X_n = \phi$ for $m \neq n$), $x_{ri}^{\alpha} \in X_r$ である. 計量は Euclid 計量を採用するため $g_{\alpha\beta} = g^{\alpha\beta}$ であり B 次元単位行列である.

Gap 統計量を用いることで, 類似度に基づく最適なクラスタリング数が式 (1) を最大化する k として与えられる.

$$(\text{最適ノード数}) = \max_k Gap(k). \tag{3}$$

これはランダムに生成したデータの場合と比べて, 実際の重み行列のデータから類似度の大きいノード群が発見される場合に非自明な最適ノード数を取ることを意味している.

3. データ概要

本研究では, 実際のビジネスで用いられている不動産ポータルサイトのデータを使用して解析を行う. 具体的なデータの内容を表 1 に示す.

表 1: 使用データ概要

対象	全国の不動産物件情報へのアクセスログ
期間	2013/1/1~2013/12/31
変数	目的変数：江東区江戸川区のモデルルームに来場したか否かのフラグ {0,1} 説明変数：ユーザが閲覧した物件の所在地のフラグ変数 {0,1}. 市区郡レベルで分割された全 376 変数
件数	72,815 件 (データから一部抽出したもの) =50,352 件 (学習)+22,463 件 (テスト)

目的変数として江東区江戸川区という特定のエリアのモデルルームへの来場フラグを設定しており, その他のエリアへ来場があったデータは除外している. これは特定のエリアに対象を絞ることで判別問題をシンプルなものにすることに加え, データの特徴を際立たせることを狙いとしている.

4. Deep Learning の設定

本研究では, Deep Learning を実行するためのライブラリとして Pylearn2 [Goodfellow 13] を用いる. Deep Learning を実行するに当たって特に重要となる設定事項を表 2 に示す. 表 2 に記載のないモデルパラメタや学習パラメタに関しては基本的には Pylearn2 の初期設定値を用いている.

表 2: Deep Learning の実行に際した設定概要

事前学習法	Stacked denoising Autoencoder
中間層層数	2
中間層ノード数	[第一層, 第二層] = [30,30]
学習率	事前学習：0.1, 教師付学習：0.1
epoch 数	事前学習：10, 教師付学習：10
活性化関数	シグモイド関数 $(1 + e^{-x})^{-1}$

事前学習において中間層第一層と第二層で独立に設定できるパラメタに関しては, ノード数以外は簡単のため共通のパラメタを用いている. 活性化関数でシグモイド関数を用いているのは, ノードの出力値を $(0, 1)$ の非負にして重み行列の重要度の解釈を容易にするためである.

5. 解析結果

本解析では, まず表 2 のように設定した状態で判別分析を実施し, その後中間層ノード数最適化と判別分析実施のセットを

表 3: 中間層ノード数が [30,30] の場合の結果

第二層	15(0.095)		
第一層	14(0.59)	9(0.54)	20(0.35)
変数	江東区 (1.4)	板橋区 (0.91)	中野区 (0.96)
	江戸川区 (1.2)	千代田区 (0.42)	杉並区 (0.74)
	台東区 (0.72)	北区 (0.41)	新宿区 (0.72)
	墨田区 (0.65)	江戸川区 (0.33)	台東区 (0.43)
	川口市 (0.62)	江東区 (0.27)	渋谷区 (0.37)

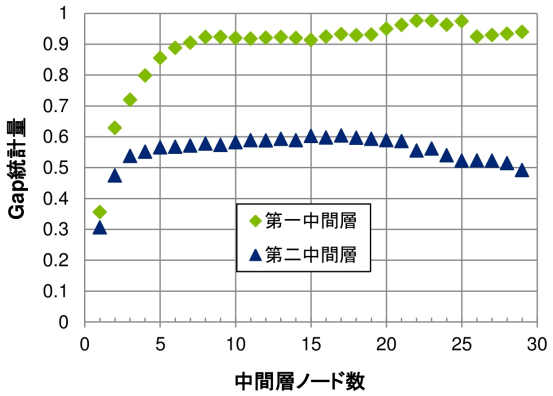

図 2: 中間層ノード数が [30,30] の場合の Gap 統計量の数値計算結果.

表 4: 中間層ノード数が [23,17] の場合の結果

第二層	17(0.13)		
第一層	5(0.55)	1(0.51)	23(0.48)
変数	杉並区 (0.98)	世田谷区 (0.68)	江東区 (2.1)
	中野区 (0.88)	川崎市中原区 (0.47)	江戸川区 (1.1)
	川崎市中原区 (0.43)	板橋区 (0.33)	墨田区 (0.33)
	武蔵野市 (0.36)	戸田市 (0.30)	台東区 (0.29)
	三鷹市 (0.22)	川口市 (0.27)	世田谷区 (0.28)

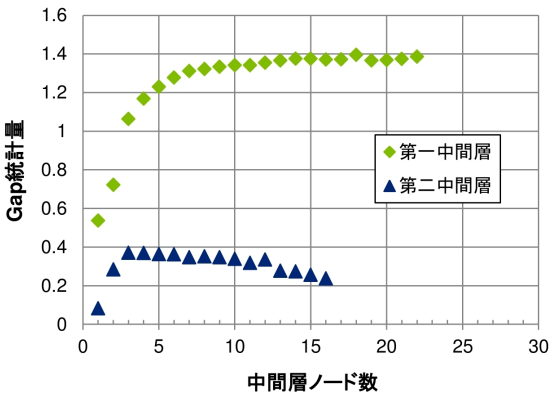

図 3: 中間層ノード数が [23,17] の場合の Gap 統計量の数値計算結果.

二回繰り返した. その結果, 中間層ノード数は [30,30] → [23,17] → [18,3] と推移し, それにつれて中間層ノードと強く結びつく説明変数や判別精度が変化する様子が観測された. 以下, それぞれの場合での詳細を記す.

5.1 中間層ノード数が [30,30] の場合

判別分析の学習により得られる重み行列から重要度の高い中間層ノードを特定し, それらのノードに強く結びつく説明変数の組を明らかにした結果が表 3 である. ここでは出力ノードに対する寄与が最も大きい第二層のノードに注目し, それに対して寄与の大きい第一層のノードを三個, さらにそれらに対して結びつきの強い説明変数を五個ずつ抽出した. 表 3 の第二層, 第一層における数字はノード番号を表しているが, これはただのラベルであり特別な意味は有していない. () 内の数字は重要度で, これは重み行列成分の値として与えられるため実数値を取る. 重要度はスケールに本質的な意味はなく層別に見たときの相対的な大きさに意味がある量である.

結果は解釈が与えやすいものとなっており, 例えば第一層ノード番号 14 にグルーピングされる説明変数は川口市を除いて江東区江戸川区付近のエリアであり, 第一層ノード番号 20 にグルーピングされる説明変数は台東区以外は西側かつ比較的価格の高いエリアである. 今回のパラメタセットの場合, 中間層が抽出する特徴は不動産物件閲覧の際の典型的なエリアであり, 納得性が高いものであった.

中間層ノード数を最適化するために Gap 統計量を計算した結果が図 2 である. 縦軸が Gap 統計量となっているが, このスケール自体には本質的な意味はなく, 層別に見たときの相対的な大きさが重要な量である.

図 2 から, どちらの場合もなだらかではあるが非自明な中間層ノード数で Gap 統計量が最大値を取ることが確認できる. 第一中間層に関してはノード数が 20 付近から値が大きくなってノード数 23 で最大となり, 第二中間層に関してはノード数が中央値付近でやや値が大きくなっていてノード数 17 で最大となる. この結果, 類似度に基づく最適ノード数は [23,17] であると結論づけられる.

5.2 中間層ノード数が [23,17] の場合

上述の結果を受けて再度判別分析を行い, 重要ノードに対する説明変数のグルーピングの結果をまとめたものが表 4 である. 第一層ノード番号 23 にグルーピングされるものは世田谷区を除いて先程と同様に江東区江戸川区付近であるが, 第一層ノード番号 5 では西側, 第一層ノード番号 1 では北側に 23 区をまたぐ説明変数のグルーピングを形成している. この結果は, 中間層ノード数が少なくなることで一つ一つの中間層ノードにより多くの説明変数が複雑に影響を及ぼすことになり, その結果目的変数との直接的な共起が少なくとも説明変数同士の共起によりグルーピングされるようになったことを示唆している.

中間層ノード数を最適化するために Gap 統計量を計算した結果が図 3 である. 第一中間層はノード数と共に増加していくがノード数 18 で局所的に最大値を取り, 第二中間層はノード数 3 で最大値を取った後減少していく様子が見て取れる. この結果, 類似度に基づく最適ノード数は [18,3] であると結論づけられる.

5.3 中間層ノード数が [18,3] の場合

先の結果得られた中間層ノード数 [18,3] で同様の解析を行った結果が表 5 である. 第一層ノード番号 3 に関しては東京の西側エリアがグルーピングされているが, それ以外では京都や兵庫・大阪など明らかに江東区江戸川区とは関係の薄い遠隔地が抽出された. 実際に元データを調べてみると, これらの関西エリアの不動産物件を閲覧しているユーザが江東区江戸川区のモデルルームに来場している場合ははほとんどない. したがって, 中間層ノード数が少なすぎると目的変数とほとんど関係のない説明変数までが強く混ざりあった特徴が抽出され, それゆ

表 5: 中間層ノード数が [18,3] の場合の結果

第二層	1(-1.1)		
第一層	3(2.6)	6(0.91)	8(0.64)
変数	世田谷区 (1.4)	京都市中京区 (1.0)	西宮市 (0.74)
	杉並区 (0.48)	京都市下京区 (0.95)	神戸市 (0.72)
	目黒区 (0.28)	京都市上京区 (0.42)	豊中市 (0.67)
	武蔵野市 (0.27)	渋谷区 (0.40)	神戸市 (0.49)
	三鷹市 (0.26)	豊中市 (0.32)	吹田市 (0.48)

え人間の理解や解釈とは乖離のある結果を生じやすくなると考えられる.

5.4 精度検証

精度指標として, 学習したモデルに予測データを投入し, 出力ノードの値を確信度スコアとするゲインチャートに対する Area Under the Curve(AUC) を採用する. 得られるゲインチャートの例と他モデルとの比較として, 図 4 に中間層ノード数が [30,30] の場合とバギングされた決定木 (CHAID アルゴリズム) の場合の結果を記載した. モデルの優劣を比較するのが目的ではないため, 今回は精度を高めるためのパラメタ調整は特に行っていない.

本研究では, 類似度に基づいて中間層ノード数を最適化した結果, モデルの判別精度にどのような影響があるかを調べた. 中間層ノード数の初期値として [30,30],[40,40],[50,50] の場合にノード数最適化をそれぞれ二回行って AUC の値の推移を調べたものが表 6 である. ノード数の最適化を施すことで, モデルの判別性能に一定の向上が見られた. これは類似度によるノード数最適化が精度を上げる一つの指針となり得ることを示唆している. しかしながら, これは一般的な性質ではなく中間層ノード数の初期値やモデル・学習パラメタ, Gap 統計量の乱数依存性などに強く依存するものであることに注意されたい.

表 6: AUC の結果

中間層ノード数	[30,30]	[23,17]	[18,3]
AUC	0.8669	**0.8766**	0.8761
中間層ノード数	[40,40]	[31,40]	[26,40]
AUC	0.8675	0.8707	**0.8718**
中間層ノード数	[50,50]	[39,43]	[34,35]
AUC	0.8722	0.8727	**0.8734**

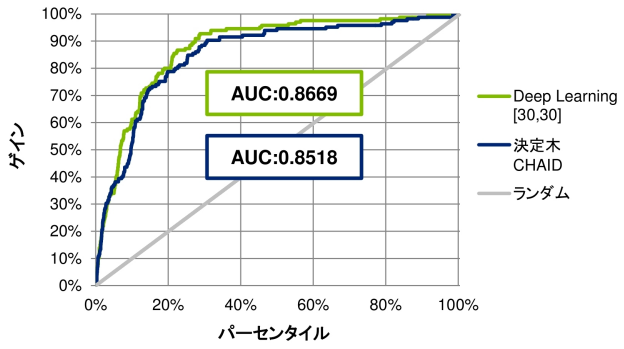

図 4: ゲインチャートの例

6. 結論と今後の展望

本研究では, Deep Learning の中間層情報を活用する試みとして, 学習後に得られる重み行列からノードの重要度や類似度を定義し, 重要度に基づいた出力結果に影響が大きい説明変数の組の特定や類似度に基づいた中間層ノード数の最適化を行う手法を提案した.

この手法を不動産ポータルのデータを対象とした二値判別分析の場合に適用した. 結果として, 重要度の高い中間層ノードが目的変数近辺のエリアや不動産物件閲覧の際の典型的なエリアをグルーピングすることが示された. また, 最適化に伴うノード数の減少につれて, そのグルーピングが目的変数とは直接的な共起が少ないエリアを含むようになる様子が観測された. 他方で, モデルの判別性能に関してはノード数最適化により上昇する傾向が得られた. Deep Learning が獲得する特徴表現を理解し利用するためには, 中間層情報をどのように取り扱うかということに加えて, 解釈可能性と識別精度のトレードオフという観点も必要になることを示唆している.

本研究では中間層ノードの類似度の定義として Euclid 距離に基づくベクトル間距離を導入したが, どのような距離が Deep Learning の中間層の活用に適したものなのかを調べることは意義深いものである. 分散を考慮したマハラノビス距離や重み行列成分を二値化した後のスペクトラルクラスタリングなど, 様々な候補を適用し中間層の把握と活用の可能性を調べていくことが今後の一つの方向性である.

今回は中間層情報自体を調べその意味を解釈することを主たる結果として提示したが, 今後は中間層情報を活用した説明力のあるモデルの構築や中間層が抽出する特徴量によるデータクラスタリングなど, 応用に即したものを研究していくことに重きを置いていく.

参考文献

[Le 12] Le, Q.V., Ranzato, M.A., Monga, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y.: Building High-level Features Using Large Scale Unsupervised Learning. In *ICML*, 2012.

[Goodfellow 13] Goodfellow, I.J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., Bengio,. Y.: Pylern2: a machine learning research library. *arXiv preprint arXiv:1308.4214,* 2013.

[Tibshirani 01] Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a dataset via the Gap statistic. J. R. Stat. Soc. B 63 (2001)411.