2C1-OS-06a-2

ヒトの尿データへの DeepLearning 適用による肺がん判定の試行と考察

Consideration and trial of lung cancer determination by DeepLearning application to human urine data

門出 康孝[*1] Yasutaka Monde
清水 徹[*1] Toru Shimizu
黒田 忠広[*1] Tadahiro Kuroda

[*1] 慶應義塾大学大学院理工学研究科
Faculty of Science and Technology, Keio University

We tried to apply a DeepLearnig to diagnose the lung cancer from a gas chromatography mass spectrometry data of human urine. The mother data consists of 28 healthy people and 39 lung cancer patient urine data sets. Each data set has 394 pieces of peak value as a feature. We applied unsupervised and supervised learning to four-layer neural network (NN) to the 57 data sets out of 67, remaining five healthy and five lung cancer data sets for testing the learning result. We got 90% accuracy of the diagnosis. We also analyzed robustness of the learning method by changing initial value, learning rate and the number of learning times.

1. はじめに

DeepLearning は従来のニューラルネットワーク(NN)を多層構造にすることで高い表現力を獲得する機械学習の手法である。これまで NN の多層化には膨大な計算量が発生し、現実的な計算時間で学習を終えることができなかった。しかし近年、GPU などのコンピュータ能力の発展によりこの問題が解決し、高い表現力を獲得した DeepLearning は画像認識、音声認識などの分野において従来手法を大幅に上回る高い精度を上げている。しかし、上記以外の分野への適用例はまだ少なく、どの分野で DeepLearning が有効かは未知数である。

DeepLearning の適用が少ない例として人間のバイタルデータがあげられる。バイタルデータといっても様々だが、今回その中でも尿に注目した。がん患者の尿には特定の化学物質が多く含まれることが知られている。尿はガスクロマトグラフィー質量分析法によって中に含まれる成分を分析、データ化ができ、そのデータを用いて DeepLearning による肺がんの検出ができるのではないかと考えた。肺がんの検出方法としては胸部 X 線や喀痰検査などが一般的であるが、それぞれの検出感度は胸部 X 線が 80%、喀痰検査が 40%前後となっており確実に検出できるとは言えない。

本研究では DeepLearning を尿の GC-MS データに適用し、高い肺がん検出率を達成することを目標とし、高検出率を達成するための前処理、正規化方法について試行を重ね、検討を行った。健常者 28 人・肺がん 39 人、1 人当たり 394 特徴点の質量分析値を母データとし学習用には健常者 23 人・肺がん 34 人の計 57 人を使用し学習を実施。テスト用には健常 5 人・肺がん 5 人の計 10 人を使用し肺がんの判定を行った。結果、90%の精度を得ることができた。

2. DeepLearning

2.1 Autoencoder を用いた教師なし学習

Autoencoder は次元の可逆圧縮を行うことで特徴の抽出を行う教師なし学習手法の一つである。本研究では教師あり学習の前の教師なし学習としてこの Autoencoder を使用した。

連絡先:門出 康孝, 慶應義塾大学大学院理工学研究科, 神奈川県横浜市港北区日吉 3-14-1, 045-566-1432, monde@kuroda.elec.keio.ac.jp

Autoencoder は図 1 のように入力 x を y へと情報の圧縮を行い、その後 z へと復元する。x・y・z はそれぞれ入力層・隠れ層・出力層と呼ばれ、この入力層の x と出力層の z が同じ値になるよう各層間の重みを調整していくのが Autoencoder である。

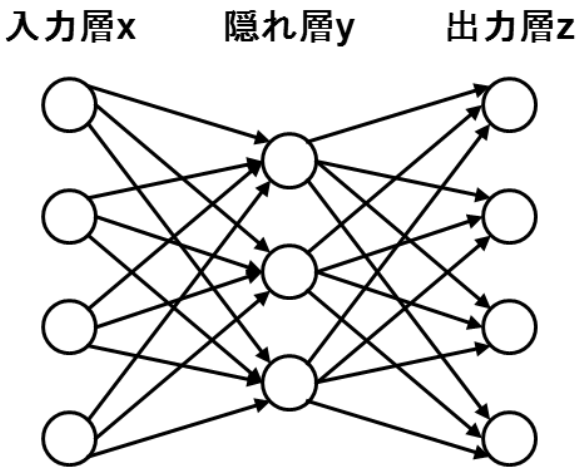

図 1 Autoencoder

各層間の計算は以下の通りである。

$$y_j = \sigma(W_{ji}x_i + b_j) \quad (1)$$

この時、x は d 次元のベクトル、W は $d \times d'$の行列、y は d'次元のベクトル、b は y に対するバイアスである。また、σはシグモイド関数

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (2)$$

次に復元部はの計算は以下の(3)の通りである。

$$z_i = \sigma(W'_{ij}y_i + b'_i) \quad (3)$$

圧縮部と同様に、y は d'次元のベクトル、W'は $d' \times d$ の行列、z は d 次元のベクトル、b'は z に対するバイアスである。以上のように入力を各層に渡り伝搬させ、入力 x と出力 z が同じ値になるようパラメータ W・W'・b・b'の調整をしていく。ここで、W と W' については tied weight と呼ばれる W の転置行列を使う方法があるが、今回は使用しなかった。

次にパラメータの更新について述べる。Autoencoder は入力と出力の値が同じになるよう調整する。そのため誤差は

$(x-z)$で表現できる。この誤差を先ほどと逆方向に伝搬させ、各パラメータの調整を行うことで学習をする。

2.2 Stacked Autoencoder を用いた教師あり学習

次に Stacked Autoencoder について述べる。Stacked Autoencoder は Autoencoder を多段に積層させ、最後に教師あり学習を行い、ネットワーク全体を調整する DeepLearning の 1 手法である。Autoencoder を積み重ねることで後段の深い層ほどより入力の特徴を抽象化した表現が得られると考えられている。Stacked Autoencoder の学習には Autoencoder を一層目から順々に適用する方法をとる。

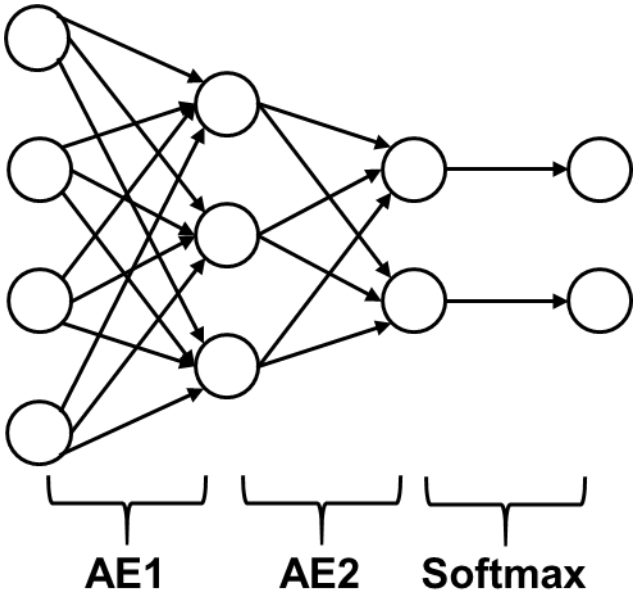

図 2 Stacked Autoencoder

図 2 は Stacked Autoencoder の概略図である。まず、AE1 の部分で式(1)の W を確定させる。その後、AE2 で AE1 の出力を入力として使用し再度 AE2 の W を確定させる。このように順々に教師なし学習を行う。最終段では Softmax 関数

$$f(x_i) = \frac{exp(x_i)}{\Sigma_{j=1}^{n} exp(x_j)}, i = 1, \dots n \qquad (4)$$

を使用し各クラスのユニット出力を計算する。教師なし学習が終了した後は教師あり学習で再度全体のパラメータを調整する。教師あり学習では最終段の Softmax 関数と教師となるクラスの値を比較し誤差を計算し、逆伝搬させ各パラメータの調整を行う。以上の計算を何度も繰り返すことで学習が完了する。もしより深いネットワークを構成したい場合は AE の数を増やせばよい。

3. DeepLearning を用いた肺がん判定

3.1 使用したデータ

本研究では DeepLearning をヒト尿データに使用し肺がんの判定を行った。使用するヒト尿データは GC-MS によって混合された有機物の組成がデータ化され保存されたものである。GC-MS によって得られた尿データは質量・保持時間・強度の 3 次元データとなっている。ヒト尿データは 394 個のピーク値を持ち今回は各ピーク値の強度を DeepLearning の入力とした。使用したヒト尿データは健常 28 人・肺がん 39 人の計 67 人分あり、DeepLearning の学習用には健常者 23 人・肺がん 34 人の計 57 人、性能評価のためのテスト用に健常者 5 人・肺がん 5 人の計 10 人を使用した。また、DeepLearning の手法としては先述の Stacked Autoencoder を使用した。

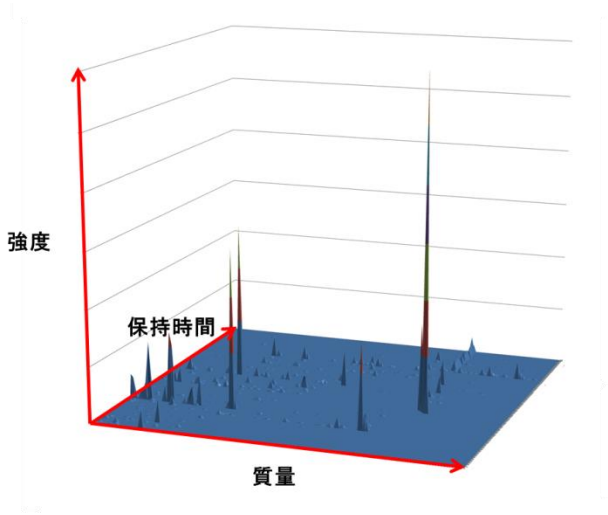

図 3 ヒト尿データ

3.2 入力データの正規化方法

GC-MS によって得られる強度データは正の整数となっている。この強度データを 0~1 の範囲へと正規化を行った。その際、複数のパターンを試し、比較を行った。

(1) 対数変換の有無

GC-MS によって生成された強度データは図 3 のように非常に大きい値を持っているものあるが、ほとんどは小さいピーク値となって表れている。0~1 の範囲に正規化する際、単純に最も大きいものが 1 となるように全データを割り算すると小さいピーク値がつぶれてしまい、その特徴が失われてしまう可能性があると考えた。そのため、まず全体の対数をとりその後最大値が 1 になるよう調整を行い実験した。比較のために対数変換なしの場合も実験を行った。図 4 に対数変換による分布の変化を示す。

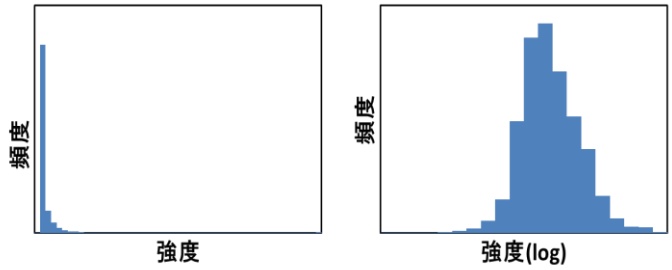

図 4 対数変化による分布の変化

(2) 最大値の取り方

GC-MS によって生成された尿のデータは個人の体調等によって濃い・薄いがあるのではないかと仮定した。濃度によって尿に含まれる有機物の量は変化するが中に含まれる割合は変化しないと考えると 0~1 の範囲に正規化する際、個人ごとに 394 個あるピーク値の最大値を 1 となるようデータを割り算すればよい。

他の方法として 394 個あるピーク値ごとに正規化する方法が考えられる。各ピーク値において有機物の有無によって値が変わるとすると最も物質が多く含まれる人のピーク値を 1 としてそれに対しどの程度物質が含まれているかを割り算して計算すればよい。以上の 2 通りを最大値の取り方として実験し比較を行った。

3.3 ネットワーク構成と学習率・学習回数

ネットワーク構成と学習率・学習回数は判定結果を鑑みて複数の値の試行を行った。

3.4 実験の流れ

最後にこれまでの実験の流れを図 5 と以下の作業 1-6 に示す。

作業 1 ： 被験者の尿を集める。

作業 2 ： 収集した尿を GC-MS によって 3 次元データ化を行う。

作業 3 ： 3 次元尿データの正規化を行う。

作業 4 ： 正規化したデータを DeepLearning に入力し学習を行う。

作業 5 ： 判定結果を確認し,正規化方法・ネットワーク構成・学習率・学習回数の変更を行う。

作業 6 ： 作業 3-6 を繰り返し最適パラメータの探索を実施する。

4. DeepLearning による肺がん判定結果

4.1 正規化方法と対数変換の効果

表 1 に正規化方法と対数変換の有無を変えて学習・判定をした際の結果を示す。以上の結果より正規化方法については個人ごとにピーク値の最大を 1 として正規化を行うことが効果的であることがわかった。また、対数変換の有無については効果が各正規化方法によってばらばらであるため、ネットワーク構成・学習率・学習回数を変えて学習を行う際も引き続き両方のパターンで実施した。

表 1 正規化方法と学習率による精度の変化

学習データ		57個	
テストデータ		10個	
ネットワーク構成		394-200-100-2	
学習率		教師なし0.1 教師あり0.5	
学習回数		教師なし300回 教師あり300回	
正規化方法		個人ごと	ピークごと
対数変換	なし	8/11	6/10
	あり	9/10	4/10

4.2 ネットワーク構成と学習率・学習回数

ネットワーク構成と学習率・学習回数を変化させ学習を行った。その結果を表 2 に示す。なお正規化方法については 4.1 の結果より個人ごとに正規化する方法が有効であると考えピークごとの正規化方法は実施しなかった。結果、学習率が教師なし学習 0.1 教師あり学習 0.1、学習回数が教師なし 300 回教師あり 300 回、正規化方法個人ごと、対数変換あり、4 層ネットワーク構成の条件で肺がん識別率 90%を達成した。以上より、人間のバイタルデータ、特に尿の GC-MS データに対する DeepLearning の有効性が示せたとともに従来の胸部 X 線や喀痰検査以上の精度を達成することができた。

5. まとめと今後の展望

今回、DeepLearning を用いて肺がん判定率 90%を達成し、胸部 X 線・喀痰検査を上回る精度を達成したが、尿を GC-MS にかけ、そのデータを解析するには費用も時間も非常にかかる。より手軽で高精度な肺がん検査手法を考えるにあたって、今回のようにたくさんのピーク値のデータから解析するのではなく肺がんの原因となるいくつかの物質に注目し解析することが重要になる。注目する物質が少なくなれば今回のように GC-MS を使うのではなく、小型のセンサでセンシングすることが可能になる。原因物質を見つけるための方法として DeepLearning の重みの強さを解析することで、どの入力が肺がんの判定に対し重要かを見つけることができるのではないかと考えており今後はこちらの検討を行っていく。また、平行して、さらに安定的に肺がんの

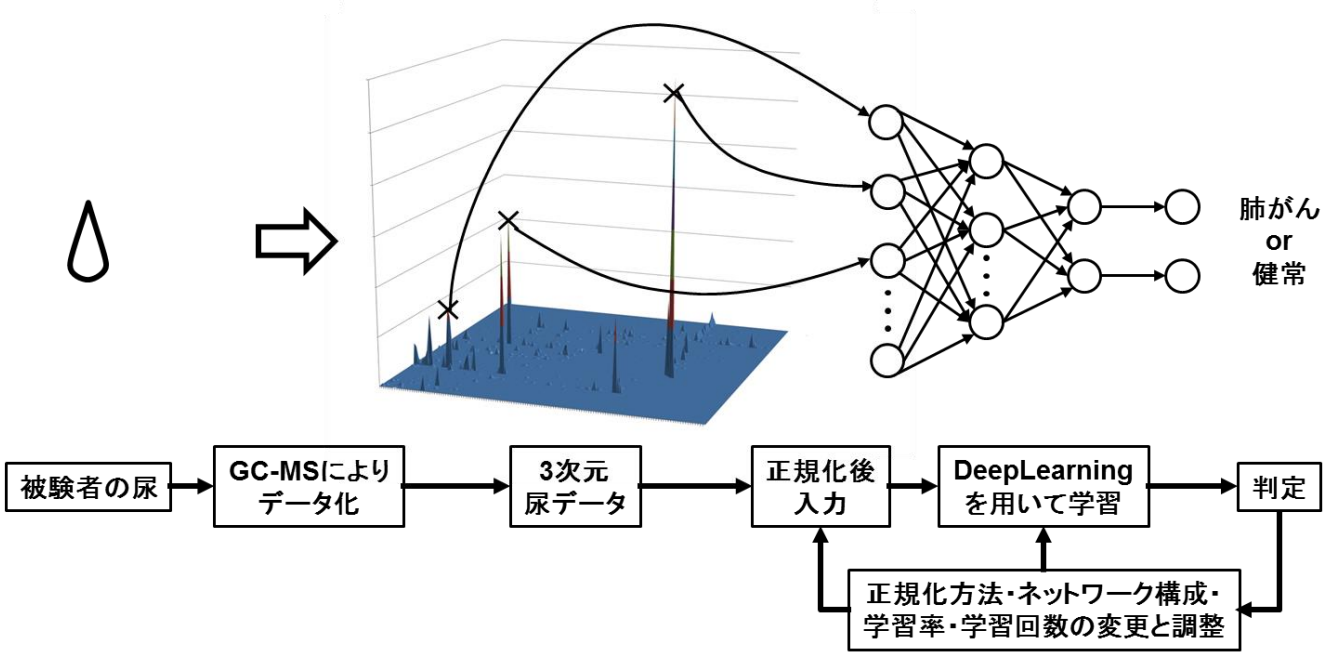

図 5 実験の流れ

表 2 ネットワーク構成と学習率・学習回数

学習サンプル		57個	57個	57個	57個	57個
テストサンプル		10個	10個	10個	10個	10個
ネットワーク構成		394-200-100-2	394-200-100-2	394-200-100-2	394-200-100-2	394-100-50-2
学習率		教師なし0.1 教師あり0.1	教師なし0.1 教師あり0.5	教師なし0.1 教師あり0.9	教師なし0.1 教師あり0.5	教師なし0.1 教師あり0.5
学習回数		教師なし300回 教師あり300回	教師なし300回 教師あり300回	教師なし300回 教師あり300回	教師なし100回 教師あり100回	教師なし300回 教師あり300回
正規化方法		個人ごと	個人ごと	個人ごと	個人ごと	個人ごと
対数変換	なし	7/10	8/10	5/10	8/10	6/10
	あり	9/10	4/10	4/10	4/10	6/10

判定を行えるネットワークを見つけるために交差検証を実施し結果の妥当性向上を目指す予定である。

参考文献

[Krizhevsky 2012] Alex Krizhevsky ,et al.: ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25, MIT Press ,2012.

[Hinton 2006] G.Hinton and R. R. Salakhutdinov : Reducing the dimensionality of data with neural networks, Science, Vol. 313. no. 5786, pp. 504 - 507 , 2006.

[Hinton 2012] G. Hinton et al. : Deep Neural Networks for Acoustic Modeling in Speech Recognition , IEEE Signal Processing Magazine, 2012.

[Liang 2014] M. Liang et al. : Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach , IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014.

[岡谷 2013] 岡谷 真樹 他: コンピュータビジョン最先端ガイド 6, アドコムメディア , 2013.