G-010

スパースモデリングによる積層自己符号化器の情報圧縮機構の明確化
A Sparse Modeling Approach to Clarifying the Mechanism of Information Compression in Stacked Autoencoders

石川眞澄†‡
Masumi Ishikawa

1. はじめに

近年の深層学習の進展は著しく、囲碁の AlphaGo はチャンピオンを凌駕し、認識能力は人をも凌ぐと言われている。従来型の機械学習では人が特徴量を与えるのに対し、深層学習はデータに基づいて特徴量を自動生成できるので、機械学習におけるブレークスルーと言われている。ただ獲得された特徴量は必ずしも明確ではなく、判断基準も人には理解できないことから、ブラックボックスと言わざるを得ないことは深層学習の深刻な問題点である。

この問題点を解決しようとする研究活動として DARPA の Explainable AI、産総研の理解できる AI を始めとして、企業においてもさまざまな研究がなされている。データ適合とコンパクトモデルを指向するスパースモデリングは有望なアプローチと考えられ、深層学習のみならずデータに基づく数理モデル全般で有望視されている[1]。

筆者は L1 ノルム（ラプラス正則化）を提案した[2][3][4]。データ適合とコンパクトモデルという利点に加えて、得られたモデル構造も入出力間の関係を理解するのに重要であると指摘した。離散入出力の場合には、この副産物としてデータに内在する規則を発見することを可能にした[5]。

L1 ノルムだけではコンパクトでかつ正確なモデルを獲得するには十分ではない。ここで正確なモデルとは、訓練データおよびテストデータに対して出力自乗誤差が小さいことを意味する。一般に L1 ノルムはモデル適合度を劣化させるので、これを改善する方策として、選択的 L1 ノルムを提案した[2][3][4]。

本稿では、積層自己符号化器を３層の自己符号化器に分割し、スパースモデリングにより各自己符号化器を学習し、これらを積み重ねたものを初期値とし、５層の積層自己符号化器をスパースモデリングにより学習する。学習結果がスパースであれば、入力層から出力層への経路を調べることにより情報圧縮機構を解明できる。この経路は大別して恒等写像と疑似恒等写像がある。これら２種類の写像と情報圧縮誤差の関係を明らかにしたい。

情報圧縮機構の解明を複雑にするのは、一つの入力素子が複数の隠れ素子と結合するという冗長表現の存在である。この存在の有無を統計的検定により判定する方法を提案した[6]のでこれを適用し、獲得された積層自己符号化器を評価する。

2. スパースモデリングについて

近年、スパースモデリングの深層学習への適用例として、スパース自己符号化器に関する多くの研究がなされている。Jiang ら[7]は平均自乗誤差評価（MSE）と隠れ層ニューロン活性度に対する L1 ノルムの組み合わせを提案した。Hosseini-Asl ら[8]、Ali ら[9]は、MSE、結合重みに対する L2 ノルム、KL 情報量を組み合わせ用いる方法を提案した。Guo ら[10]、Zeng ら[11]は、MSE、KL 情報量の組み合わせを提案した。Goodfellow ら[12]は、MSE、結合重みに対する L1 ノルムを提案した。また自己符号化器ではないが、Yang ら[13]、Cogswell ら[14]は、過剰学習を避けるため共分散のうち非対角成分の最小化という正則化項を提案した。

ただ、知る限りではこれらで得られたモデル構造や隠れ層ニューロンの意味づけなどは示されていないので、スパースモデリングの深層学習への適用は必ずしも十分とは言えない。またさまざまな正則化項が用いられているが、これらの有効性を適切に評価できているわけではない。筆者は、モデル適合度及びスパース度からなる二つの評価軸からなるパレート最適性の概念を正則化項の有効性の評価基準とすることを提案した[15]。

3. スパースモデリングの定式化

L 個の隠れ層及び出力層は tanh 関数を用いる。学習時の評価関数としては出力自乗誤差(SSE)を採用する。入力を $x_i^{(n)}, i=1,...,I; n=1,...,N$ 、隠れ層出力を $h_{lj}^{(n)}, l=1,...,L; j=1,...,J_l; n=1,...,N$ 、出力層出力を $y_k^{(n)}, k=1,...,K; n=1,...,N$ 、目標出力を $t_k^{(n)}, k=1,...,K; n=1,...,N$ 、データ数を N とする。この時、各隠れ層及び出力層の出力値及び出力自乗誤差は下記で表される。

$$h_{1j}^{(n)} = \tanh(\sum_{i=1}^{I} W_{ij}^{(l)} x_i^{(n)} + b_j^{(l)}),\ j=1,...,J_1; n=1,...,N \quad (1)$$

$$h_{lj}^{(n)} = \tanh(\sum_{i=1}^{I} W_{ij}^{(l)} h_{(l-1),i}^{(n)} + b_j^{(l)}),\quad j=1,...,J_l; l=2,...,L; n=1,...,N \quad (2)$$

$$y_k^{(n)} = \tanh(\sum_{j=1}^{J} W_{jk}^{(L+1)} h_{lj}^{(n)} + b_k^{o}), k=1,...,K; n=1,...,N \quad (3)$$

$$SSE = \sum_{n=1}^{N}\sum_{k=1}^{K}(y_k^{(n)} - t_k^{(n)})^2 \quad (4)$$

ただし、$W_{ij}^{(l)}$ は下から l 番目の隠れ層と下層との間の結合重み行列、$b_j^{(l)}$ は下から l 番目の隠れ層の j 番目のニューロンのバイアス値である。

L1 ノルムは式(5)で与えられ、λ_1 はそのウエイトである。

$$l_1 = \lambda_1(\sum_{i=1}^{I}\sum_{j=1}^{J_l}\sum_{l=1}^{L}|W_{ij}^{(l)}| + \sum_{j=1}^{J_l}\sum_{l=1}^{L}|b_j^{(l)}| + \sum_{k=1}^{K}|b_k^{o}|) \quad (5)$$

† （一財）ファジィシステム研究所 Fuzzy Logic Systems Institute
‡ 九州工業大学 Kyushu Institute of Technology

L1 ノルムは出力自乗誤差を大きくするので、L1 ノルム学習後、式(6)の選択的 L1 ノルムを適用する。ここで L1 ノルム学習後に結合重み値が閾値 θ_w 以上のものは必要性があるので残存していると考え、結合重み値が閾値以下の結合のみ L1 ノルムによる減衰の対象とする。これにより出力自乗誤差が更に減少する。

$$l_{1s} = \lambda_1 \Big(\sum_{i,j,l,|W_{ij}^{(l)}|<\theta_w} |W_{ij}^{(l)}| + \sum_{j,l,|b_j^{(l)}|<\theta_w} |b_j^{(l)}| + \sum_{k,|b_k^o|<\theta_w} |b_k^o| \Big) \quad (6)$$

4. 計算機実験の手順

積層自己符号化器の学習を以下の手順で行う。

1) 与えられたデータを入力及び出力目標として用い、隠れ層上に圧縮情報が得られるよう第１段自己符号化器を学習する。この際、式(4)＋式(5)を最小化する。（これを bpl1 学習と略称）なおここでは{入力層、隠れ層、出力層}にそれぞれ{21，8，21}個の素子を用いる。

2) 第１段自己符号化器の隠れ層出力を第２段自己符号化器の入力および出力目標として学習する。学習時の評価関数は第１段の時と同じ bpl1 学習である。なおここでは{入力層、隠れ層、出力層}にそれぞれ{8，5，8}個の素子を用いる。

3) ２個の学習済み自己符号化器を積み重ね積層自己符号化器の初期値とし、これまでと同様 bpl1 学習を行う。ここでは入力層から出力層までの素子数がそれぞれ{21，8，5，8，21}個である。

4) このままでは L1 ノルムのため出力自乗誤差が大きいので、さらに小さくするため式(4)＋式(6)を評価関数とする。すなわち L1 ノルムの代わりに選択的 L1 ノルムを用いる。（bpsl1 学習と略称）

積層自己符号化器の隠れ層ニューロン間に相関があるか否かを、相関係数の母平均が零であるという帰無仮説の統計的検定により判定する。

相関係数の仮設検定は、標本相関係数を r とおくと、$t = \dfrac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$ が自由度 n-2 の t 分布に従い、サンプル数 n が十分大きい場合、$z = \dfrac{\sqrt{n-3}}{2}\ln\left(\dfrac{1+r}{1-r}\right)$ が正規分布 N(0、 1)で近似できることから、有意水準（ここでは 5%）に対応する帰無仮説の棄却域を設定できる。

5. 計算機実験結果

ここでは UCI の胎児心拍数陣痛計データを用いた[16][17]。サンプル数は 2126 であり、21 属性を表１に示す。ある属性値が他属性値より桁違いに大きいなど存在範囲にばらつきが見られるので、学習を安定的かつ効率的に行うため適切な正規化を行った。なお、以下の計算機実験においては、ランダムに選択した 1915 個のサンプルを訓練データ、211 個のサンプルをテストデータとする。本来は胎児の状態を{正常、疑わしい、異常}の３クラスに分類するクラス分類課題であるが、ここではクラス情報は無視し、自己符号化器として 21 個の属性値の再現をタスクと考える。訓練データに対する共分散行列を図１に、21 属性を表１に示す。

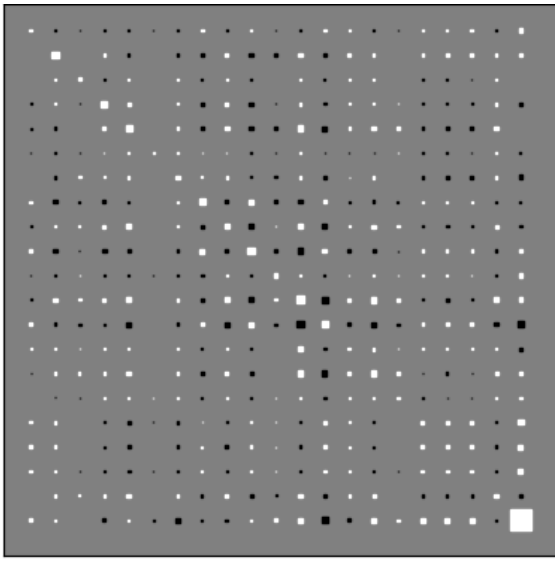

図１．胎児心拍数陣痛計訓練データの共分散行列。白は正の共分散、黒は負の共分散を表し、共分散値は面積に比例する。

表１．胎児心拍数陣痛計データ。”#”は個数を表す。

No.	label	attributes
1	LB	FHR baseline (beats per minute)
2	AC	# of accelerations per second
3	FM	# of fetal movements per second
4	UC	# of uterine contractions per second
5	DL	# of light decelerations per second
6	DS	# of severe decelerations per second
7	DP	# of prolongued decelerations per second
8	ASTV	percentage of time with abnormal short term variability
9	MSTV	mean value of short term variability
10	ALTV	percentage of time with abnormal long term variability
11	MLTV	mean value of long term variability
12	Width	width of FHR histogram
13	Min	minimum of FHR histogram
14	Max	Maximum of FHR histogram
15	Nmax	# of histogram peaks
16	Nzeros	# of histogram zeros
17	Mode	histogram mode
18	Mean	histogram mean
19	Median	histogram median
20	Variance	histogram variance
21	Tendency	histogram tendency

実は UCI の赤ワインデータを用いた積層自己符号化器の学習を過去に実施した[6]。これは 12 属性であったのに対し今回は 21 属性でありより複雑なデータを用いて情報圧縮機構を明らかにするという目的に加えて、新たに正則化項の評価を行うとともに、恒等写像と疑似恒等写像の性能比較を定量的に実施するという目的を持つものである。

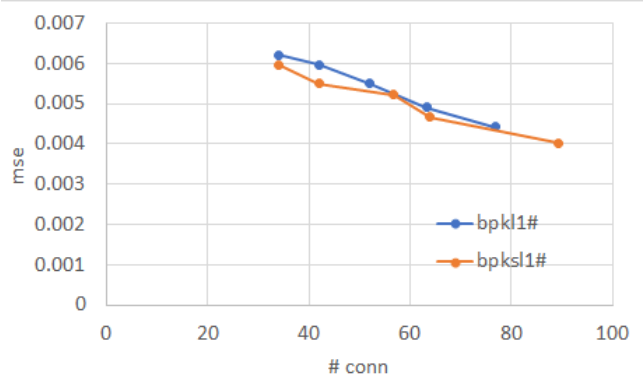

図２．L1 ノルムと選択的 L1 ノルムの性能比較。縦軸は平均出力自乗誤差(MSE)、横軸は結合数である。

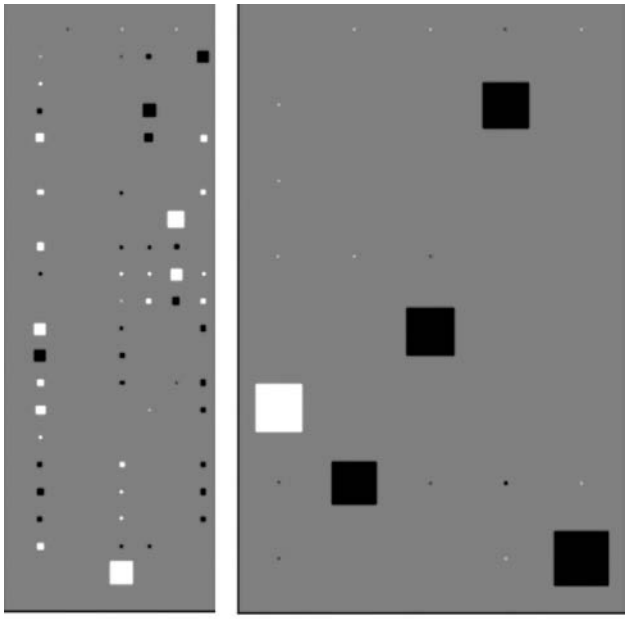

図３．積層自己符号化器の結合重み行列。左が入力層から第１隠れ層までの結合重み、右が第１隠れ層から第２隠れ層までの結合重み。第２隠れ層から第３隠れ層までの結合重みは右の転置、第３隠れ層から出力層までの結合重みは左の転置であるので、図は割愛する。

L1 ノルムと選択的 L1 ノルムを用いた学習、すなわち bpl1 学習と bpsl1 学習の性能比較を図２に示す。学習率は 0.2、式(5)の L1 ノルムのウエイトとして図の右から順に {0.00005、 0.00007、 0.0001、 0.00015、 0.0002}、学習手順の中で 1)2)3)の bpl1 学習は 20 万回、これに続く 4)の bpsl1 学習における結合重みの閾値は 0.1 を用い、10 万回の学習を行った。L1 ノルムのウエイトが大きくなるに従い、MSE 値は増大し結合数は減少する。

２節で述べたように、データ適合度及びスパース度という二つの評価軸からなるパレート最適性の概念を導入して正則化項の有効性の評価を行った。ここでスパース度としては結合数を用いる。図２で原点により近い折れ線の方が MSE も結合数も少なくパレート最適の意味で優れている、この意味で bpl1 学習よりも bpsl1 学習の方が優れていることが分かる。この差はそれほど大きくないが、クラス分類課題[15]ではもっと大きな差がある。bpl1 学習と bpsl1 学習の差がそれほど大きくないのは連続値出力目標を有する自己符号化タスク特有の問題と考えている。

図３は積層自己符号化器の４個の結合重み行列のうち、下半分の２個の結合重みを示す。左が 21x8、右が 8x5 の行列である。図４は５層からなる積層自己符号化器の構造を示す。いずれも L1 ノルムウエイトが 0.00015 の場合である。図３では結合重み値が白（正）あるいは黒（負）の面積によって表現され、図４では結合重み値がグレイ（正）あるいは黒（負）の線幅で表現されている。線幅よりも面積の方が表現能力がずっと大きいため、図３ではすべての結合重みが表現されているが、図４では絶対値が 0.1 より大きい結合重みのみが表現されている。

図３及び図４より、学習により得られた積層自己符号化器は、１個の恒等写像と４個の疑似恒等写像から構成されている。恒等写像と異なり、疑似恒等写像は複数の入力素子と結合する。入力層を *i*、隠れ層を下から順に、*1h, 2h, 3h*、出力層を *o* で表現し、その後の数字が左から何番目の素子かを表すものとする。まず、*i21→1h5→2h3→3h5→o21* という入力から出力への経路が恒等写像に相当している。２入力の疑似恒等写像として、*{i4, i5}→1h6→2h1→ 3h6→{o4, o5}*、及び*{i2, i5}→1h8→2h5→3h8→{o2, o5}*がある。３入力の疑似恒等写像として、*{i8, i10, i11}→1h7→ 2h2→3h7→{o8, o10, o11}*がある。８入力の疑似恒等写像として、*{i5, i9, i12, i13, i14, i15, i18, i20}→1h2→2h4→ 3h2→{o5, o9, o12, o13, o14, o15, o18, o20}*がある。

恒等写像は情報損失がないのに対し、疑似恒等写像は情報損失がある。ただ隠れ素子数が入力素子数よりも小さいので、全てを恒等写像で実現することは不可能である。他属性との相関が小さく分散の大きな属性が恒等写像で実現される傾向にあると考えられる。逆に他属性との相関が大きな属性群が疑似恒等写像で実現される傾向にあると考えられる。図３、図４におけるこれらの傾向を表３に示す。

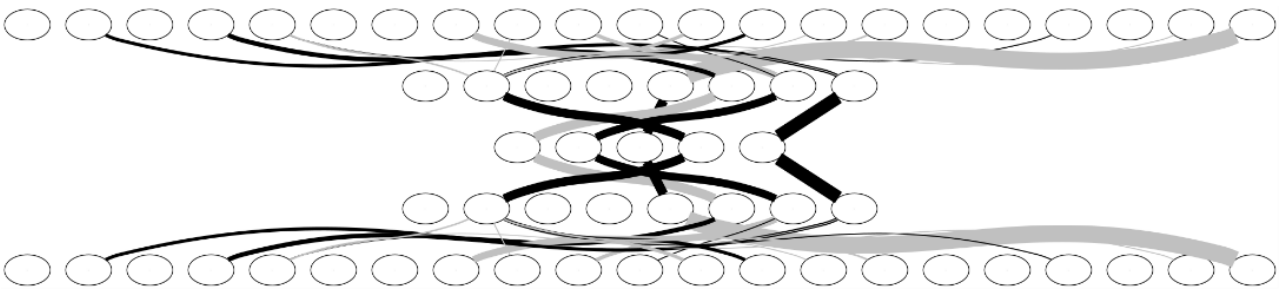

図４．学習後の積層自己符号化器構造

表２．L1 ノルムウエイト(weight)と各隠れ層における隠れ素子出力間の標本相関係数が棄却域（有意水準 5%の場合、絶対値が 0.04479 より大）に存在する百分率(%)。hidden1, hidden2, hidden3 はそれぞれ第１隠れ層、第２隠れ層、第３隠れ層を表す。

weight	hidden1	hidden2	hidden3
0.00005	24.29	6.00	21.43
0.00007	17.17	14.00	23.23
0.00010	23.08	12.00	23.08
0.00015	0.00	0.00	0.00
0.00020	0.00	0.00	0.00

表３．恒等写像と疑似恒等写像の特性比較

	恒等写像	疑似恒等写像
A データ分散	0.3709	0.0187
B 学習で説明できる部分	0.3702	0.0129
B/A 説明できる比率	0.9980	0.6927

恒等写像ではデータ分散をほぼ学習により説明できているが、疑似恒等写像では説明できない部分が３割ほど残っている。

この中で入力*"i5"*は２個の２入力疑似恒等写像、８入力疑似恒等写像に関与している。すなわち*"i5"*の情報が*{1h2, 1h6, 1h8}*という３個の第１隠れ層の素子と結びついているので冗長表現が生じる。これら隠れ素子間に相関がありそうに思われるが、この間の結合重みがそれぞれ{0.25, -0.21, 0.14} と小さいため、表２の下から２段目に示すように相関係数が零であるという帰無仮説が棄却されず、無相関と考えて差し支えないことが分かる。

また、ここでは主として L1 ノルムのウエイトとして 0.00015 を用いたが、表２よりウエイトが 0.0001 以下では結合数が増え、相関係数が零であるという帰無仮説が棄却される。

6. おわりに

本稿では、UCI の胎児心拍数陣痛計データを用い、積層自己符号化器を複数の自己符号化器で分割学習し、これらを積み重ねたものをさらに学習する。学習時にスパースモデリングを用いて積層自己符号化器をスパース化した。スパースな５層の積層自己符号化器の入力層から出力層への経路を調べることにより、情報圧縮機構を解明できた。この経路は大別して恒等写像と疑似恒等写像に分けられる。

学習後の積層自己符号化器の構造から、１個の恒等写像と４個の疑似恒等写像により情報圧縮を行っていることが判明した。恒等写像は情報損失が無いが、入力素子数よりも隠れ素子数がずっと小さいので、多くの場合は情報損失のある疑似恒等写像を用いざるを得ない。実験結果によると、恒等写像には殆ど情報損失がなく（データ分散のうち殆どを説明できる）、疑似恒等写像には一定程度の情報損失がある（データ分散のうち一定部分が説明できない）ことが分かった。またデータ分散が大きな入出力素子に対して恒等写像が適用されており、全体的な情報損失を減らす働きと考えられる。

情報圧縮機構の解明を阻害するのは、一つの入力素子の情報が複数の隠れ素子と結合するという冗長表現の存在である。胎児心拍数陣痛計データを用いた学習の中でも主として示したケースで１個の入力素子が第一隠れ層の 3 個の素子と結合した結果が見られた。ただ、これらの結合重み値が小さいので隠れ素子間の相関係数が零であるという帰無仮説が統計的検定により棄却されず、無相関と考えて差し支えないという結論となった。

スパース自己符号化器においてさまざまな正則化項が用いられているが、これらの有効性を適切に評価した例はない。筆者は、モデル適合度及びスパース度という二つの評価軸からなるパレート最適性の概念を正則化項の有効性の評価基準とすることを提案し、本稿での L1 ノルムよりも L1 ノルム学習後に選択的 L1 ノルムを適用するとさらに性能が向上することを示した。

なお、本研究は JSPS 科研費 JP18K11487 の助成を受けたものである。

参考文献

[1] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Statist. Soc. Series B, vol.58, 1996.
[2] M. Ishikawa, A structural learning algorithm with forgetting of link weights, International Joint Conference on Neural Networks, 1989.
[3] M. Ishikawa, A structural learning algorithm with forgetting of link weights, Technical Report TR-90-7, Electrotechnical Laboratory, pp.1-17, 1990.
[4] M. Ishikawa, Structural learning with forgetting, Neural Networks, Vol.9, No.3, pp.509-521, 1996.
[5] M. Ishikawa, Rule extraction by successive regularization, Neural Networks, Vol.13, No.10, pp.1171-1183, 2000.
[6] 石川眞澄、積層自己符号化器における冗長表現およびブラックボックスの抑制、電子情報通信学会ニューロコンピューティング研究会、2019.12.
[7] X. Jiang et al., A Novel Sparse Auto-Encoder for Deep Unsupervised Learning, 6th International Conference on Advanced Computational Intelligence, pp.256-261, 2013.
[8] E. Hosseini-Asl et al., Deep Learning of Part-Based Representation of Data Using Sparse Autoencoders with Nonnegativity Constraints, IEEE Trans. NNLS, Vol.27, Issue 12, pp.2486-2498, 2016.
[9] A. Ali et al., Automatic modulation classification using deep learning based on sparse autoencoders with nonnegativity constraints, IEEE Signal Processing Letters, VOL. 24, NO. 11, pp.1626-1630, 2017.
[10] Y. Guo et al., Deformable MR prostate segmentation via deep feature learning and sparse patch matching, IEEE Trans Med Imaging. 35(4), pp.1077-1089, 2016.
[11] N. Zeng et Al., Facial expression recognition via learning deep sparse auto-encoders, Neurocomputing, vol.273, pp.643-649, 2018.
[12] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, pp.496-498 The MIT Press, 2016.
[13] J. Yang et al., Learning structured and nonredundant representations with deep neural networks, Pattern Recognition, vol.86, pp.224-235, 2019.
[14] M. Cogswell et al., Reducing overfitting in deep networks by decorrelating representations, ICLR 2016.
[15] 石川眞澄, 層毎貪欲学習および各種正則化項によるクラス分類深層ネットワークのスパース化, 電子情報通信学会ニューロコンピューティング研究会, 2020.3.
[16] UCI Machine Learning Repository, Cardiotocography Data Set, https://archive.ics.uci.edu/ml/datasets/cardiotocography 2010.
[17] Ayres de Campos et al., SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms, J Maternal Fetal Medicine, vol. 9, pp.311-318, 2000.