

平成 18 年度 「自主課題研究」

ニューラルネットワークによる花粉の飛散量の予測

情報システム工学科

3 年 018 小林 透

0408060411

目次

- 研究の背景
- 研究課題
- 研究内容
 - ニューラルネットワーク（パーセプトロンを用いた）について理解する
 - プログラムを使用し、簡単な学習・テストを行う
 - 花粉の飛散量を予測するためのデータを集める
 - 3のデータを使用し学習・テストをする
- まとめ
- 自主課題研究を通して得たもの

・ 研究の背景

現在、ニューラルネットワークといえばパターン認識やデータマイニングに応用されているが、その原理や構造がどうなっているかについて、今回の自主課題研究を通して実際にそれらを学び、実装したプログラムを使い理解を深めることを目的とする。

・ 研究課題

最終的な目標として「ニューラルネットワークにより、花粉の飛散量を予測する」ことを研究課題として設定する。

研究内容は

1. ニューラルネットワーク（パーセプトロンを用いた）について理解する
2. プログラムを使用し、簡単な学習・テストを行う
3. 花粉の飛散量を予測するためのデータを集める
4. 3のデータを使用し学習・テストをする

という手順で行うものとする。

・ 研究内容

1. ニューラルネットワークについて理解する

まず、ニューラルネットワークとは、脳機能を担っている神経細胞（ニューロン）を計算機上のシミュレーションによって表現しようとする数学モデルである。

代表的な人口ニューラルネットワーク

- ・ フィードフォワードニューラルネット

最も早くに考案された単純な構造の人口ニューラルネットワークモデルであり、ネットワーク間にループする結合を持たず、入力から出力までの結合が単一方向のみに信号が伝播するもの

- ・ リカレントニューラルネット

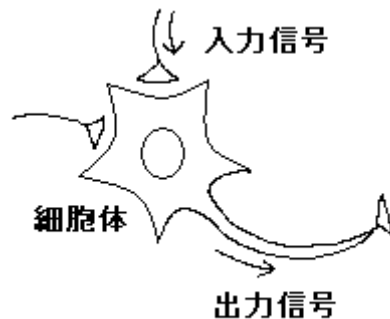
全ての人口ニューロン間で双方向に信号が伝播するモデル。

- ・ 確立ニューラルネット

乱数による確率的な動作を導入した人口ニューラルネットワークモデル。

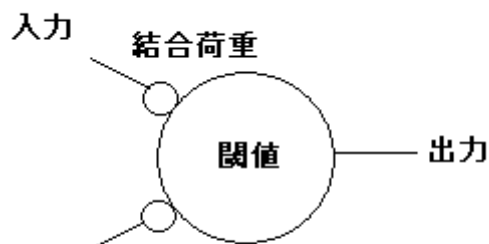
◆ ニューロン

ニューラルネットワークは脳細胞のニューロンを数学モデル化したものであるのでニューロンについて簡単に説明する。



上の図はニューロンを簡単に表したものであり、入力信号が入ってくると、ニューロンそれぞれにある伝達効率によって、細胞体に入る電圧が変化する。これが閾値を越えると、膜電位（ニューロンと細胞外液との電位差で、約 -70mV ）が 0V を超え、出力信号が発信される。

◆ ニューロンモデル



ニューロンを数学モデル化したものが上の図のニューロンモデルであり、伝達効率を結合荷重という部分で任意の実数値で入力に重みを掛けるという方法で実現している。

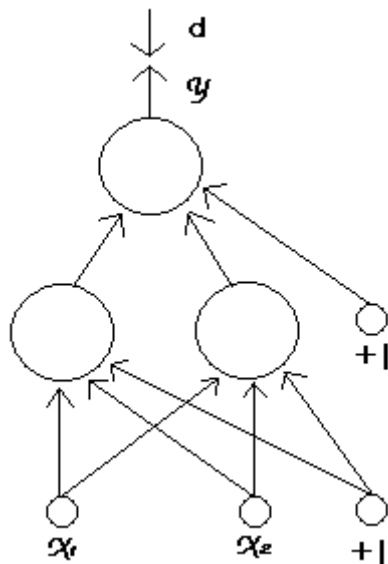
出力と実際に得たい結果の差の2乗（2乗平均誤差）が小さくなれば学習されていると考えられる。

2. プログラムを使用し、簡単な学習・テストを行う

花粉の飛散量を予測する前に、E X - O R を用いて学習とテストを行う

EX-OR. 真理値表

x_1	x_2	d
0	0	0
0	1	1
1	0	1
1	1	0



x_1 x_2 を入力とし、 d をターゲット

入力層のユニット数 2

隠れ層のユニット数 2

出力の数 1

学習率 0.1

繰り返し回数 10000

として学習させる。

テスト結果は次の通り

rep : 繰り返し回数

mse : 2乗平均誤差

output : 出力

```
rep*****mse*****output***x1*****x2*****
0          0.092526 0.608361 0.000000 0.000000
0          0.042964 0.614351 0.000000 1.000000
0          0.038971 0.619024 1.000000 0.000000
0          0.102825 0.633684 1.000000 1.000000
1          0.091141 0.603791 0.000000 0.000000
1          0.043807 0.609563 0.000000 1.000000
1          0.039937 0.614260 1.000000 0.000000
1          0.101354 0.628835 1.000000 1.000000
*****
9998       0.000248 0.062998 0.000000 0.000000
9998       0.000269 0.063624 1.000000 1.000000
9998       0.000233 0.941229 0.000000 1.000000
9998       0.000229 0.941436 1.000000 0.000000
9999       0.000214 0.941501 1.000000 0.000000
9999       0.000267 0.063717 1.000000 1.000000
9999       0.000265 0.062990 0.000000 0.000000
9999       0.000232 0.941286 0.000000 1.000000
rep*****mse*****output***x1*****x2*****
```

mse は十分小さくなっており、学習出来ている。この重みで似た値でテストしてみる

x ₁	x ₂	Output
0.10 $\div$ 0	0.10 $\div$ 0	0.155084 $\div$ 0
0.20 $\div$ 0	0.90 $\div$ 1	0.925091 $\div$ 1
0.82 $\div$ 1	0.02 $\div$ 0	0.946140 $\div$ 1
0.89 $\div$ 1	0.92 $\div$ 1	0.141873 $\div$ 0

これより、プログラムがニューラルネットワークとして機能していることが分かる。

3.花粉の飛散量を予測するためのデータを集める

花粉の飛散量がニューラルネットワークで予測できる
ものだとすれば、予測に有効な要因として考えられる

- ・ 平均気温
- ・ 降水量
- ・ 日照時間
- ・ 夏日（25℃を超える日）
- ・ 降水の無い日（降水が 1mm 未満の日）
- ・ 毎年の花粉の飛散量

のデータを集める。

「平均気温・降水量・日照時間・夏日・降水の無い日」は気象庁のホームページに掲載されている気象統計情報のデータを用いる。

参照： 気象庁 <http://www.jma.go.jp/jma/index.html>

昨日までのデータ <http://www.data.kishou.go.jp/etrn/index.html>

「毎年の花粉の飛散量」については全国各地の統計をしている機関がなく、主に地方の民間の団体が行い、データ管理しているところが多いことが分かった。

今回、データは石川県医師会のホームページに掲載されているデータ 13 年分を使用した。

参照： 石川県医師会 <http://www.ishikawa.med.or.jp/>

ここで、スギ花粉についての予備知識として

1. 前年の夏にスギの花芽がつくられる
2. スギの雄花は秋に最も成長する
3. 秋から冬にかけて一旦休眠状態になる
4. 冬の寒さが刺激になり開花の準備を始める
5. 2 月中旬ごろから飛散し始める
6. スギ花粉は 2, 3 周期で飛散が増減する
7. 飛散するのは主に晴れの日で、雨の日はほぼ飛散しない

といった仕組みになっている。

実際に役に立つものを作るとすれば、前年のデータだけで次の年の花粉の飛散量を求めることが出来る方が好ましいので、入力データとして

- ・ 平均気温
- ・ 降水量
- ・ 日照時間

この 3 つについてはスギ花粉の性質を踏まえた上で「7 月～11 月」のデータを使用することにする。ここで、12 月の気象はスギ花粉の「開花時期」を決定する因子に当たるので、今回は 11 月までを使用する。

ここまでのことをまとめると

- 春のスギ花粉の飛散量は前年の夏から秋の気候に左右される
- 花粉の飛散には 2, 3 年の周期がある
- 実際の飛散量は降水が無い日のみのものである

これらを予測結果に反映するために

→前年の 7 月～11 月のデータを用いる

→その年と過去 3 年の花粉の飛散量を入力する

→飛散量は、飛散した年の 2 月～5 月の降水の無い日との比で求める。

ここで、**周期の入力**について詳しく述べると

(例えば花粉が 3 周期だとすると

入力			出力
少ない	少ない	多い	少ない
多い	少ない	少ない	多い
少ない	多い	少ない	少ない

というように周期を学習させようということである。

・正規化

ここまでのデータを用い、学習・テストを行うが、扱う値が0～1なので、これらのデータを正規化しなければならない。**それぞれの値を最大値で割る**という方法で正規化したところ、年ごとにあまり変化のしない気温データが、ほとんど1に近い値となってしまう、これではニューラルネットワークが区別できないということで、次のように正規化した

7・8月の気温 → -15°C

9・10月の気温 → -10°C

11月の気温 → -5°C

した値を正規化した。これにより、0～1に値が分布する幅が広がった。

4. 学習・テストをする

以上を踏まえて学習・テストを行う

13 年分の花粉の飛散量のデータがあるが、周期を入力として入れるため、過去の飛散量が必要となるので、実際にデータとして使えるのは 10 年分である。

具体的に学習とテストは

0 0	0 1	0 2	0 3	0 4	0 5	年
			学習 ←	テスト		
		学習 ←	テスト			

このように 04 年までを学習させ、残りをテスト、03 年までを学習・残りをテストといったように行う。

■10 年分・花粉の周期あり

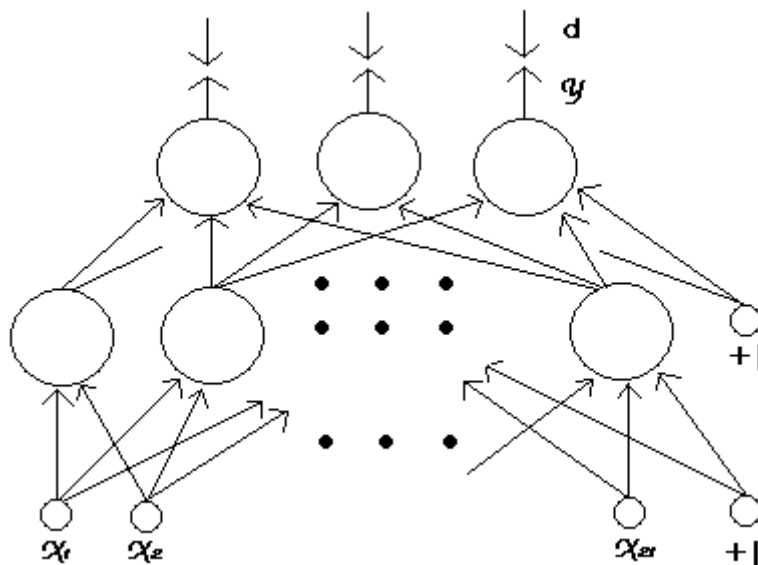
入力層のユニット数 21

隠れ層のユニット数 9

出力の数 3

学習率 0.1

繰り返し回数 10000



ここでの出力の数3というのは、多い・普通・少ない の3段階で出力させるということで、多い=100、普通=010、少ない=001、とする。

…として学習させた。学習させた入力に対しては100%の正答率を得ることができたが、未知のデータに対する結果は次の通り

ニューラルネットワークの出力			実際の値	年度	合否
0.074195	0.032118	<u>0.596621</u>	100	2004	×
0.026015	0.003476	<u>0.958901</u>	010	2005	×
0.002657	0.111021	<u>0.953195</u>	001	2003	○
0.019541	0.042044	<u>0.900183</u>	100	2004	×
0.022298	0.017288	<u>0.954022</u>	010	2005	×
0.001768	<u>0.672454</u>	<u>0.712418</u>	100	2002	×
0.001456	0.372030	<u>0.914334</u>	001	2003	○
0.078435	0.043295	<u>0.749555</u>	100	2004	×
0.005000	0.085777	<u>0.954841</u>	010	2005	×
<u>0.940814</u>	0.086631	0.008010	010	2001	×
0.015114	0.155585	<u>0.819595</u>	100	2002	×
0.009034	0.046929	<u>0.961823</u>	001	2003	○
0.104363	0.032393	<u>0.728964</u>	100	2004	×
0.025531	0.015080	<u>0.958613</u>	010	2005	×

学習したデータに対する正答率 100%

正答率 21.4%

考察

下線が引いてあるものは大きな値のものである。この結果を見てみると、全体的に001（花粉の量が少ない）に結果が偏っていることが分かる。正答率で考えると約20%しかない。しかし、この正答率も2003年が花粉の飛散量が少ないための結果であるので、単純に「少ない」になるように重みが偏ってしまっていると考えられる。

また、学習に使ったデータは花粉が多い年、普通の年、少ない年を含んでおり、普通の年が少し少ないが、花粉の飛散量が少ない年ばかりを学習しているためではないと考えられる。

そこで、もっと単純な条件で学習が出来ているかを確認するために **花粉の周期**を入力しないで学習をさせてみた結果が次の通り

ニューラルネットワークの出力			実際の値	年度	合否
<u>0.286286</u>	0.037721	<u>0.285196</u>	1 0 0	2004	△
0.012989	0.005811	<u>0.983988</u>	0 1 0	2005	×
0.001144	0.273921	<u>0.798739</u>	0 0 1	2003	○
<u>0.355946</u>	0.046833	0.159800	1 0 0	2004	○
0.019645	0.009084	<u>0.973892</u>	0 1 0	2005	×
0.009029	0.008812	<u>0.969846</u>	1 0 0	2002	×
0.004484	<u>0.685856</u>	0.134570	0 0 1	2003	×
<u>0.651320</u>	0.037640	0.118343	1 0 0	2004	○
0.013246	0.014973	<u>0.939274</u>	0 1 0	2005	×
<u>0.462198</u>	<u>0.454480</u>	0.011915	0 1 0	2001	△
0.007744	0.036539	<u>0.897487</u>	1 0 0	2002	×
0.000860	0.336631	<u>0.804070</u>	0 0 1	2003	○
<u>0.928034</u>	0.019984	0.050195	1 0 0	2004	○
0.007144	0.007187	<u>0.979205</u>	0 1 0	2005	○

学習したデータに対する正答率 100%

完全に一致している正答率 42.9%

おいしいものも一致しているとする正答率 57.1%

考察

花粉の周期を入力したときと比べて明らかに正答率が上がっていることが分かる。

このことより、周期のデータに結果を悪くする要因があると考えられる。

具体的に、正規化する際に花粉の飛散量のデータは **最大値で割る** という方法で正規化してあるのだが、この際に **異常に花粉が大量に飛散した年** があり、その数値で正規化したために、他の値が小さくなりすぎたことが結果に大きく影響していることが考えられる。

前回の結果を元に、正規化の方法を変える

最大値で正規化 → 最大値の半分未満の値で正規化

結果は次の通り

ニューラルネットワークの出力			実際の値	年度	合否
0.004066	0.108532	<u>0.952716</u>	1 0 0	2004	×
0.012759	0.009925	<u>0.985993</u>	0 1 0	2005	×
0.005376	0.133501	<u>0.956722</u>	0 0 1	2003	○
0.003788	0.331317	<u>0.907490</u>	1 0 0	2004	×
0.013958	0.023333	<u>0.980886</u>	0 1 0	2005	×
0.001683	<u>0.820800</u>	0.470133	1 0 0	2002	×
0.000993	0.297601	<u>0.983025</u>	0 0 1	2003	○
0.001749	0.149386	<u>0.817956</u>	1 0 0	2004	×
0.001841	0.149386	<u>0.986742</u>	0 1 0	2005	×
<u>0.893504</u>	0.184911	0.010969	0 1 0	2001	×
0.004993	0.423419	<u>0.705788</u>	1 0 0	2002	×
0.001429	0.051453	<u>0.985568</u>	0 0 1	2003	○
0.005960	0.395619	<u>0.695013</u>	1 0 0	2004	×
0.005325	0.013325	<u>0.986479</u>	0 1 0	2005	×

学習したデータに対する正答率 100%

正答率 21.4%

考察

結果としては正規化の方法を変更する前と同じ結果になった。内容は、より正解に近くなるところもあれば、より遠くなるところもあるので、決して改善されたとは言えない。このことから、花粉の周期を学習させる点において、学習させるデータにニューラルネットワークが周期を見出せていないと考えられる。これは、花粉の周期が 2 年と 3 年が混合して存在し、それらをニューラルネットワークが学習できないためだと考えられる。

これまでの結果より、花粉の周期を取り入れることは難しく、結果を悪くする要因として強く働いてしまうということが分かったので、花粉の周期のデータを入れないことにする。また、これにより、過去 3 年分の花粉の飛散量のデータが入力に使えることになるので、13 年分を学習とテストに使用することとする。

周期を考えないので学習とテストの方法を

「12 年を学習・1 年をテスト」

という方法に変え、これを 13 パターン行うことにする

結果は次の通り

ニューラルネットワークの出力			実際の値	年度	合否
<u>0.096806</u>	0.004153	<u>0.901534</u>	0 1 0	2005	×
<u>0.968403</u>	0.026749	0.004486	1 0 0	2004	○
0.000161	<u>0.898527</u>	0.550593	0 0 1	2003	×
0.002211	<u>0.987650</u>	0.034667	1 0 0	2002	×
<u>0.925432</u>	0.119093	0.006084	0 1 0	2001	×
<u>0.998737</u>	0.001405	0.003132	1 0 0	2000	○
0.030145	0.008928	<u>0.692159</u>	1 0 0	1999	×
0.037046	<u>0.942166</u>	0.014884	0 0 1	1998	×
<u>0.991880</u>	0.003637	0.002926	0 0 1	1997	×
<u>0.949913</u>	0.046484	0.005612	0 1 0	1996	×
0.000036	0.318232	<u>0.960913</u>	0 0 1	1995	○
0.317927	<u>0.858668</u>	0.002523	1 0 0	1994	×
0.000432	0.002130	<u>0.997785</u>	0 0 1	1993	○

学習したデータに対する正答率 100%

正答率 30.7%

考察

前回よりも低い結果となった。花粉の量が普通（0 1 0）に対しての正答率が 0%であり、普通を多いと捉えているようである。多い（1 0 0）に対する正答率は 40%であり、少ない（0 0 1）に対する正答率も 40%であった。また、多い（1 0 0）と少ない（0 0 1）が間違えるときは両方とも普通（0 1 0）と間違えることが多いようだ。

まとめ

ニューラルネットワークで学習が出来ているかどうかは、学習したデータに対する正答率から学習するところまではうまく出来ていると考えられる。

しかし、未知のデータから花粉の飛散量を予測するにあたって、花粉の周期を組み込むことにより、逆に結果に悪い影響を与えているということが分かった。また、学習データを増やした結果が、学習データが少ない場合よりも悪い結果となったので、このことから、①入力に使用したデータ以外にも花粉の飛散量を予測するために必要なデータがある、または、②その年毎にあまり関係性がない、ということになってしまうのだが、10年間のデータで学習とテストを行っていたときに正答率 50%ほどの結果が得られていたので、①が原因であると考ええる。

また、それとは別に、多い・普通・少ない を区別するためのデータが少なすぎるものが問題だと考えられ、途中で断念した花粉の周期というものも実測されたデータから結果がずれる原因だとも考えられる。

改善策

単純に、実在するスギの木は様々な気象の状態や周期によってその年の花粉の量というのが左右されていると考えられるので、学習させるデータを増やすということが最も有力な改善策だと考えられる。

また、気象データなども、もっと細かくしたほうがよいかもしれないし、一日ごとの降水量の変化が影響を与えているかもしれないと考ええると、入力のパターンを変えることが改善に繋がるとも考えられる。

問題

気象のデータは気象庁がかなり昔のものまで管理・保存してくれてあるのだが、肝心の花粉のデータは基本的に民間の団体が近年始めたことであり、あまりデータが残っていないというのが一番の問題である。

自主課題研究を通して得たもの

今回、自主課題研究によって、その名前は聞いたことがあったが、その具体的な内容・仕組みまでは理解していなかった「ニューラルネットワーク」について、簡単だが触れることができ、自分にとってとても有意義なものになった。

時間が設けられ、その時間の間それを知る者が教えてくれるという普通の授業とは違い、自主的にインターネットや文献から自分が必要とする情報を探し出し、自分のものとするこの難しさや、それ故に得られる満足感を味わうことができた。

特に、週に 1 回の報告の際に、自分では思いもしなかったことをアドバイスしてもらえたので、どんな問題も、やはり一人で考えるよりももっと大勢で考えたほうが良いということも実感した。

また、集めたデータ、得られた結果を初めて見るひとが見やすいようにまとめるということの難しさを改めて感じた。