The Comparison of Generalized Additive Model with Artificial Hierarchical Neural Network in the Analysis of Pharmaceutical Data

Tatsuya, TAKAGI\textsuperscript{a,b}, Emiko KUROKAWA\textsuperscript{a}, Kouji MIYATA\textsuperscript{a}, Kousuke OKAMOTO\textsuperscript{a}, Yuko TANAKA\textsuperscript{a}, Ken KUROKAWA\textsuperscript{b}, Teruo YASUNAGA\textsuperscript{b}

\textsuperscript{a} Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamadaoka, Suita, Osaka 565-0871 JAPAN
\textsuperscript{b} Genome Information Research Center, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871 JAPAN

(Received January 28, 2002; Accepted March 22, 2002)

Although the Generalized Additive Model (GAM) is known as a superior nonparametric regression method, there have been only a few applications, especially in the fields of chemistry and pharmaceutical sciences. GAM can be applied to nonlinear problems that can also be solved using the hierarchical Artificial Neural Network (ANN) method. In this study, GAM was compared with ANN in regression, classification, and prediction power using artificial and actual pharmaceutical data sets. The results show that GAM and ANN have similar regression/classification/prediction powers. Considering the fact that additive models simply visualize the relationship between a predictor variable and a response variable, GAM can be applied to data sets in the pharmaceutical sciences.

Key Words: non-parametric regression, generalized additive model, artificial neural network, pharacoepidemiology, metric pharmaceutical science

Introduction

The Multiple Regression Model:

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$  \hspace{1cm} (1)

has been widely used to reveal the linear relationships

*ttakagi@phs.osaka-u.ac.jp
between response and predictor variables. Here, $X_j$ indicates the $j$-th predictor variable and $y$ indicates the response variable. It is one form of the traditional linear regression model. In the field of pharmaceutical sciences, the Logistic Regression Model,

$$\log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$  \hspace{1cm} (2)

is also commonly used as a linear regression model. Here, if the data of a response variable is a binary number, $\mu$ indicates the generation probability. This model relates the binary response variable to the predictor variables via the logit link function,

$$\log\left(\frac{\mu}{1-\mu}\right)$$.

Because other link functions, such as the Cox function, are sometimes used, these linear models are generically named Generalized Linear Models (GLM). The GLM,

$$g(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$  \hspace{1cm} (3)

replaces the response variable of a multiple regression model by a specific link function, $g(\mu)$.

For more flexible fitting based on a nonlinear model, Tibshirani, et al.,1) developed the Generalized Additive Models (GAM), which replaces each linear term of the GLM by a more general function form:

$$g(\mu) = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$  \hspace{1cm} (4)

where each $f_j$ is a non-parametric function, which is estimated in a flexible way using a scatter plot smoother. The Generalized Additive Models are such flexible statistical models that they can find significant nonlinear relationships between predictors and response variables. Therefore, these models can be applied to any situation where a generalized linear model has been typically used. By using these models, one can obtain information about the relationships between predictors and response variables—relationships that cannot be revealed by ordinary regression models. Furthermore, because it is based on additive models, as shown in equation (4), GAM has the advantage that it can graphically show the linear or nonlinear relationship between response and predictor variables; this is unlike many other nonlinear methods, such as artificial neural network methods, which cannot easily show these nonlinear relationships graphically. The features of GAM, such as robustness for outliers, have been studied in detail.21 However, GAM has not been commonly used in the field of pharmaceutical sciences.

On the other hand, the Artificial Neural Network (ANN)3) is a well-known method for revealing nonlinear relationships between predictors and response variables. ANNs are modeled after the signal transduction system of the brain, and have been widely used as flexible nonlinear pattern recognition methods in medical research.4) Nevertheless, there have not been enough systematic comparisons of these two methods for nonlinear problems. In this paper, we present the features of each method. We use two data sets and more than ten kinds of predictor parameter sets to analyze medical and pharmaceutical data in order to reveal the capabilities of both methods in the medical and pharmaceutical fields.

### Computation

**Constructing the Artificial data**

We generated ideal artificial data sets so that they could validate the predictability of each method. Because this artificial data was generated from random numbers, which follow a uniform distribution, neither the GAM nor ANN methods should have any prediction power for the data sets. Nevertheless, due to the flexibilities of both methods, some predictability can occur. However, predictability has no value for regression analyses performed on random numbers. In other words, regression or discriminant analyses are not expected to show predictability for data sets that are generated from random numbers. Hence, a method demonstrating less predictability for such data is more valuable.

The artificial data consisted of four predictors and one response variable. All variables for the regression analyses were generated from random numbers that follow a uniform distribution. For the discriminant analyses, two of the predictors were generated from random numbers that follow a uniform distribution, and the other two predictors were generated from successive categories of random numbers, ranging from 1 to 8. The response variable was generated from binary random numbers (0 or 1). There were 200 observations for the artificial data set labeled data-1, and 100 observations for the artificial data set labeled data-2.

### Clinical Data

**Clinical data-1:** Sugawara, et al.,5) used multiple regression analyses to determine the relationships between the molecular properties and intestinal absorption rates of some different drugs—that is, drugs with different molecular properties and absorption rates. Because they adopted only the linear regression method in revealing the linear relationships between them, we applied GAM and ANN methods for further development.

**Clinical data-2:** Breslow, et al.,6) used epidemiological data to teach biological statistics. We used an esophageal cancer data set from their many data sets to assess the relationships between some risk factors, such as alcohol and tobacco consumption, and esophageal cancer morbidity.
Computation Details

Most of the computations were carried out on a Fujitsu S4/7000 UNIX workstation at the Genome Information Research Center at Osaka University. SPSS version 9.0J (for Windows98 operating systems) and S-Plus version 4.0 (on a PC/AT compatible personal computer) were used for both the linear discriminant and multiple regression analyses. SPSS (for UNIX operation systems) at the Cybermedia Center, Osaka University, was also used. PSDD (Perceptron-type Neural Network Simulator) was used to train the artificial neural networks. The number of neurons in the hidden layer was optimized in order to satisfy the minimum prediction error condition using the external validation method. The entire artificial neural network used in this study has a three-layer structure.

Results

Artificial Data

We used the residual standard deviation and the number of classification errors as indicators of the classification power for GAM and ANN. Table 1 shows the results of regression and discriminant analyses of the artificial data using the two methods. In the results of the discriminant analyses using the ANN, the data with output signal values above 0.5 were classified as 1 (case) and the ones below 0.5 were classified as 0 (control).

Table 1. Results of Discrimination and Regression Analyses using Artificial Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Pattern1</th>
<th>Pattern2</th>
<th>Pattern3</th>
<th>Pattern4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>77/200 (38.5)</td>
<td>69/200 (34.5)</td>
<td>39/100 (39)</td>
<td>39/100 (39)</td>
</tr>
<tr>
<td>ANN</td>
<td>0.26309</td>
<td>0.30568 [10]</td>
<td>0.24164</td>
<td>0.25511 [10]</td>
</tr>
</tbody>
</table>

The numbers in () indicate percentages of classification errors and the ones in [ ] indicate the optimal numbers of hidden-layer neurons.

Clinical Data-1

We applied the two methods, GAM and ANN, to the data set of the intestinal absorption rate of the different drugs. The logarithm of the apparent absorption clearance (logP_a) was used as the response variable. The data set of 42 observations, with the various kinds of predictor variable sets, was taken from the literature. We prepared four patterns of predictor variable sets, which are shown in Table 2. Although the values of the response variable were originally expressed as quantitative data, we converted them to binary data for the logistic regression analyses.

Table 2. Predictor Variable Sets Used for the Analyses of Intestinal Absorption Rate of Drugs

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Predictor Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern1</td>
<td>( \log P_a, N_H )</td>
</tr>
<tr>
<td>Pattern2</td>
<td>( \log P_a, H_a, H_b )</td>
</tr>
<tr>
<td>Pattern3</td>
<td>( \log D_{iso}, \log D_w, N_H )</td>
</tr>
<tr>
<td>Pattern4</td>
<td>( \log D_{iso}, \log D_w, H_a, H_b )</td>
</tr>
</tbody>
</table>

\( D_{iso} \) isooctane-water partition coefficients at pH 6.0
\( D_w \) diffusion coefficients in water (37°C)
\( N_H \) hydrogen-bonding capacities
\( H_a \) hydrogen-bonding donor activities
\( H_b \) hydrogen-bonding acceptor activities
\( P_a \) permeation rate across EVA

The external validation method was adopted for validating the models of the analyses for clinical data-1. Ten observations, which were used to check the accuracy of the predictions, were randomly removed from the original data. Next, the remaining 32 observations of the data set were used for the PSDD training. Tables 3–6 show the results of the analyses for clinical data-1. In these tables, LDA is the classical linear discrimination analysis, and MRA is the multiple regression analysis. We computed the number of classification errors for the binary response variable, and the residual standard deviations for the quantitative response variable.

Table 3. Results of Discriminant Analyses using Clinical Data-1 (Binary Data)

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of discrimination errors (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern1</td>
<td>Pattern2</td>
</tr>
<tr>
<td>GAM</td>
<td>3/32 (9.38)</td>
</tr>
<tr>
<td>ANN</td>
<td>5/32 (15.6)</td>
</tr>
</tbody>
</table>

The numbers in () indicate percentages of classification errors and the ones in [ ] indicate the optimal numbers of hidden-layer neurons.
hidden-layer neurons was 6. The optimal number of hidden-layer neurons.

The numbers in [] indicate the optimal numbers of hidden-layer neurons.

Table 4. Results of Regression Analyses using Clinical Data-1 (Quantitative Data)

<table>
<thead>
<tr>
<th>Method</th>
<th>pattern1</th>
<th>pattern2</th>
<th>pattern3</th>
<th>pattern4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>0.1809</td>
<td>0.1546</td>
<td>0.2251</td>
<td>0.1760</td>
</tr>
<tr>
<td>ANN</td>
<td>0.2707</td>
<td>0.1995</td>
<td>0.3278</td>
<td>0.3053</td>
</tr>
<tr>
<td>MRA</td>
<td>0.2007</td>
<td>0.1768</td>
<td>0.2796</td>
<td>0.2441</td>
</tr>
</tbody>
</table>

The numbers in ( ) indicate percentages of classification errors.

Table 5. Prediction Results of Clinical Data-1 (Binary Data)

<table>
<thead>
<tr>
<th>Method</th>
<th>pattern1</th>
<th>pattern2</th>
<th>pattern3</th>
<th>pattern4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>4/10 (40)</td>
<td>4/10 (40)</td>
<td>3/10 (30)</td>
<td>2/10 (20)</td>
</tr>
<tr>
<td>ANN</td>
<td>2/10 (20)</td>
<td>2/10 (20)</td>
<td>4/10 (40)</td>
<td>2/10 (20)</td>
</tr>
<tr>
<td>LDA</td>
<td>5/10 (50)</td>
<td>3/10 (30)</td>
<td>6/10 (60)</td>
<td>3/10 (30)</td>
</tr>
</tbody>
</table>

The numbers in () indicate percentages of classification errors.

Table 6. Prediction Results of Clinical Data-1 (Quantitative Data)

<table>
<thead>
<tr>
<th>Method</th>
<th>pattern1</th>
<th>pattern2</th>
<th>pattern3</th>
<th>pattern4</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>0.3175</td>
<td>0.3483</td>
<td>0.3377</td>
<td>0.3333</td>
</tr>
<tr>
<td>ANN</td>
<td>0.1905</td>
<td>0.1760</td>
<td>0.2488</td>
<td>0.3040</td>
</tr>
<tr>
<td>MRA</td>
<td>0.4811</td>
<td>0.2711</td>
<td>0.3066</td>
<td>0.2531</td>
</tr>
</tbody>
</table>

Clinical Data-2

This data set used five prediction variables: age, amount of tobacco consumption per day, amount of beer consumption per day, amount of wine consumption per day, and the total amount of alcohol consumption per day.

The external validation method was applied to clinical data-2 to obtain the index of prediction power. The 581 observations, which resulted from randomly removing 97 observations from a total of 978, were used as training data. Then, the same 97 observations were used as test data. Table 7 shows the results of the analyses of clinical data-2. The optimal number of hidden-layer neurons was 6.

Table 7. Discrimination and Prediction Results of Clinical Data-2

<table>
<thead>
<tr>
<th>Method</th>
<th>Discrimination</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAM</td>
<td>151/881 (17.1)</td>
<td>54/285 (18.9)</td>
</tr>
<tr>
<td>ANN</td>
<td>203/881 (23.0)</td>
<td>30/285 (10.5)</td>
</tr>
<tr>
<td>LDA</td>
<td>200/881 (22.7)</td>
<td>88/285 (30.9)</td>
</tr>
</tbody>
</table>

The numbers in () indicate percentages of classification errors.

Discussions

Considering the properties of the artificial data prepared for this study, one can assume that the better method produces more discrimination errors. Because there was little difference between the results of the GAM and ANN methods, we concluded that the unexpected ability of classification was almost the same, at least for the data used in this study.

Because there was evidence of a relationship between the response variables of the two clinical data sets and the predictor variables, both the number of discriminant errors and residual standard deviations were expected to be small. Comparing the results of the two clinical data sets, one can conclude, at a minimum, that GAM is equal to or better than ANN in regression/classification powers for both regression and discrimination analyses, even though ANN was expected to show much better results because of its model complexity. Although GAM seems to show worse prediction results according to Table 6, it is thought that these results are due to the fact that GAM is based on a Maximum Likelihood Method, which doesn’t try to decrease the residual standard deviations. As a matter of fact, Table 5 doesn’t show any results that are worse than those for ANN. Moreover, the fact that the number of hidden-layer neurons of ANN was optimized using an external validation method must have affected the results, which showed better prediction powers for ANN compared to GAM. However, in general, the purpose of the nonparametric regression method is not thought to be to predict the response from novel data of prediction variables but, rather, to reveal the regression structures for a data set. Thus, the prediction power of the nonparametric regression method is not expected to be the same as the one of ANN. In this study, we investigated whether the prediction power of the nonparametric regression method was notably worse than ANN. As a result, the prediction powers of GAM for the clinical data sets were not thought to be notably worse than those of ANN. Instead, GAM showed a little better fitting power than ANN, as shown in Tables 3, 4, and 7.

The LDA and the MRA also show fairly good prediction power compared to both GAM and ANN. One reason is that the clinical data sets have comparatively linear relationships between the response and prediction variables. Therefore, one can conclude that such classical methods should be the first choice. Other nonlinear methods, such as the nonparametric regression methods, are preferable if the classical linear methods cannot be adopted satisfactorily. The standardized regression/discriminant coefficients for the two clinical data sets are shown as supplementary material.

Although it is generally difficult to obtain the effect of each predictor from ANN, additive models such as GAM and ACE (Alternative Conditional Expectation) are useful for obtaining those effects.
can show such information in a partial regression plot. The partial regression plots for pattern1 of clinical data-1 are shown in Figure 1 (other plots are given as supplementary material). The predictors that give an even, less varied curve have a weak effect on the response variable, and the predictor that gives a varied curve shows a strong effect on the response variable. For example, the partial regression plot of $N_H$ in Figure 1 indicates that the apparent absorption clearance is at the maximum when $N_H$ is about 3.5. This nonlinear relationship between $N_H$ and the response variable was not mentioned by Sugawara et al.\(^5\) Although it is not easy to explain what the nonlinear relationship physicochemically indicates, because the detail mechanism of the intestinal absorption of drugs is now under study, a nonlinear relationship between intestinal absorption rate and polar molecular surface areas was reported by Stenberg.\(^8\)

![Figure 1. Partial residual plots of pattern1 of clinical data-1 (binary data).](image)

In the epidemiological data set of esophageal cancer (clinical data-2), only linear relationships were detected, even though the relationship between cancer morbidity and the total consumption of alcohol has been detected in many studies. According to the GAM results of this study, a moderate nonlinear relationship between cancer morbidity and tobacco consumption is detected. This indicates that the quantity of tobacco consumption is not important for morbidity due to esophageal cancer, but whether or not one smokes is important. Therefore, these nonparametric methods enable us to detect nonlinear relationships between risk factors and link function. Furthermore, such nonlinear relationships can easily be visualized when additive models are used.

If GAM showed notably worse regression/classification/prediction powers, one could conclude that ANN should be adopted for data sets in the pharmaceutical sciences. However, because of the abovementioned facts, including the fact that GAM had similar regression/classification/prediction powers, and that additive models enable us to easily find the relationships between response and prediction variables, we have concluded that GAM is a nonlinear method that has the possibility to be quite applicable for datasets in the field of pharmaceutical sciences.

**Conclusion**

Generalized Additive Models (GAM) and ANN showed similar regression/classification/prediction powers at least for the data used in this study. GAM has the possibility to be quite applicable for data sets in the field of pharmaceutical sciences, given that the relationship between a predictor variable and a response variable is considered a key to such an analysis.

Two clinical data sets were analyzed using the two methods, GAM and ANN. By using GAM, we revealed the structure of the nonlinear relationship between the hydrogen-bonding capacities and intestinal absorption rate.

**Acknowledgement**

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 13672253, 2001.

**References and Notes**


薬学データの解析における一般化加法モデルと階層型ニューラルネットワーク法の比較

高木達也1,2, 黒川絵美子1, 宮田幸治1, 岡本晃典1, 田中優子1, 黒川顕2, 安永照雄2

1 大阪大学大学院薬学研究科、大阪府吹田市山田丘1-6
2 大阪大学遺伝情報実験センター、大阪府吹田市山田丘3-1

一般化加法モデル(GAM)は優れたノンパラメトリック回帰法の一つであるが、化学、薬学分野ではわずかな応用例しか知られていない。GAMは階層型ニューラルネットワーク法(ANN)と同様非線形問題に広く適用可能だと考えられるため、今回、GAMとANNを、人工データや実際の薬学データを用いて系統的な比較を試みた。その結果、GAMはANNとほぼ同様の、回帰、判別性能を持つことが示された。GAMは、視覚的に予測変数と応答変数間の関係を容易に補足できるなどのANNに比べて優れた点もあるため、薬学分野のデータに十分に適用可能であると考えられる。

キーワード: non-parametric regression, generalized additive model, artificial neural network, pharmacoepidemiology, metric pharmaceutical science

* ttakagi@phs.osaka-u.ac.jp