Заглавие документа

Овсянников А.В., Байда Ю.А., Лаврентьев В.С. Информационные алгоритмы обучения нейронных сетей // Труды БГТУ. Сер. физ.-мат. наук и информ. Вып.XII. 2004. С.110-113.

Авторы: Овсянников А. В., Байда Ю.А., Лаврентьев В.С.

Тема: Нейроинформатика и управление

Дата публикации: 2004

Издатель: УО Белорусский государственный технологический университет

Аннотация: В статье рассматриваются информационные алгоритмы обучения нейронной сети, когда вместо минимизации целевой функции ошибки сети в виде квадратичной формы используется минимизация функции потерь. При этом целевая функция ошибки сети в виде квадратичной формы является частным случаем функции потерь. Рассмотрен робастный подход к построению алгоритмов настройки сети для дискретного и аналогового случая. Приведенные алгоритмы доступны для моделирования в среде Matlab.

УДК 621.9

А.В. Овсянников, доцент; Ю.А. Байда, В.С. Лаврентьев, студенты

ИНФОРМАЦИОННЫЕ АЛГОРИТМЫ ОБУЧЕНИЯ НЕЙРОННЫХ СЕТЕЙ

Основные проблемы разработки и применения нейронных сетей (НС) сосредоточены в трех направлениях: выбор оптимальной структуры (топологии) сети, под конкретную задачу, определение оптимального количества нейронов использованных в сети, и определение алгоритма настройки (обучения) сети.

Остановимся на последней задаче. Стандартным и широко распространенным методом настройки синаптических весов W нейронов сети является минимизация ошибки обобщения [1,2]. При таком подходе требуемая модель G системы F не может быть основана полностью на явных правилах и формальных законах. Процесс получения G из имеющихся экспериментальных сведений о системе F рассматривается как обучение модели G поведению F в соответствии с заданными критериями. При обучении "с учителем" критерием минимизации может служить минимизация оценки ошибки E(X,W) (целевой функции) модели на обучающей выборке {X,Y}.

$$E = \sum_{x \in X} \left\| G(X,W) - Y \right\|_2 , \tag{1}$$

где G – модель системы F на некотором конечном наборе параметров X. Однако указаный выбор параметра E ограничивает область выбора алгоритма для оценки матрицы весов W. Действительно, в тривиальном случае работы НС, не состоящей из одного многовходового нейрона при ограниченном объеме обучающей выборки, в частности, при $N \le 30$, условия центральной предельной теоремы о нормализации среднего арифметического независимых одинаково распределенных величин (при конечной дисперсии) не выполняются и применение формулы (1) является не совсем корректным.

Более общим подходом, позволяющим синтезировать так же и устойчивые (робастные) алгоритмы оценки W, является информационный, который рассматривает информацию о распределении ошибки в виде плотности распределения вероятностей (ПРВ) ошибки:

$$P(E) = P_E(Y|X,W) = P_E(Y - G(X,W)) , \tag{2}$$

где $P_E(Y|X,W)$ – функции правдоподобия. Для построения информационного алгоритма обучения НС воспользуемся методом максимального правдоподобия (ММП) /3/, который основан на нахождении градиента целевой функции (2) $\mathrm{grad}_W \ln P_E(Y|X,W) = 0$. Робастный ММП состоит в использовании плотности $P_E^0(Y|X,W)$, отвечающей "наихудшему" (наименее благоприятному) для данного X распределению из заданного класса распределений вместо $P_E(Y|X,W)$, где "наихудшее" распределение определяется из условия [3] $P_E^0(Y|X,W) = \arg\min\limits_{P_E \in \Re_n} \mathrm{tr}(I)$; I – информационная матрица Фишера; $\Re_n$ – класс распределений $P_E(Y|X,W)$. Таким образом, оценка параметра W сводится к реализации алгоритма

$$W^* = \arg\max_W P_E^0(Y|X,W). \tag{3}$$

В вычислительном отношении часто удобнее рассматривать вместо P_E^0 функцию $\Im^0(E) = -\ln[P_E^0]$. Эта функция представляет собой функцию потерь, выбор которой определяет качество искомой оценки W^*. При этом оптимальной в классе $\Re_n$ функцией потерь будет та, для которой выполняется условие $D[\Im_E^0, P_E^0] \ge D[\Im_E^0, P_E]$, где D – асимптотическая корреляционная матрица погрешностей оценки W^*.

Воспользовавшись процедурой Роббинса – Монро (метод стохастической аппроксимации) в дискретной форме, получаем

$$W_{k+1}^* = W_k^* + \gamma_k \operatorname*{grad}_W \Im_E^0, \tag{4}$$

где γ_k – подходящим образом выбираемая последовательность неотрицательных чисел [3,4].

Конкретизируем уравнение (4) на примере многослойной нейронной сети. Согласно указанной выше процедуре для класса всех невырожденных распределений, когда известно только то, что оно существует: $\Re_1 = P : P_E(0) \ge \varepsilon > 0$, «наихудшее» распределение представляет собой лапласовское $P^0(E) = \frac{\lambda}{2} e^{-\lambda|E|}$, где $\lambda>0$, параметр распределения.

В этом случае алгоритм, реализующий процедуру Роббинса – Монро для рекуррентного изменения веса связи нейрона i с нейроном j для выбранного слоя НС, имеет вид

$$\omega^*_{ij,k+1} = \omega^*_{ij,k} + \gamma_k \mathrm{sign}\left[E(\omega^*_{ij,k})\right] \frac{dE(\omega^*_{ij,k})}{d\omega^*_{ij,k}}, \tag{5}$$

где

$$E(\omega^*_{ij,k}) = G_j(y_j - \sum_{i=0}^{q} \omega^*_{ij,k} x_{ij}), \tag{6}$$

G_j – в общем случае нелинейная функция характеризующая выход нейрона j, $\gamma_k = \lambda(k+1)^{-1}$, при условии, что дисперсия ошибки не превосходит величины $2/\lambda^2$.

Для класса распределений с ограниченной дисперсией $\Re_2 = \left\{ P : \int E^2 P(E) dE \le \sigma_E^2 \right\}$ имеем “наихудшее” в классе распределение, описываемое гауссовской (нормальной) ПРВ. В этом случае дискретный алгоритм (4) следующий:

$$\omega^*_{ij,k+1} = \omega^*_{ij,k} - \gamma_k E(\omega_{ij,k}) \frac{dE(\omega^*_{ij,k})}{d\omega^*_{ij,k}}, \qquad (7)$$

где $E(\omega^*_{ij,k})$ определяется формулой (6), а коэффициент $\gamma_k = (k+1)^{-1}\sigma^{-2}$, при условии, что дисперсия ошибки не превосходит величины σ_E^2.

Особый интерес представляет реализация рассмотренных выше алгоритмов в непрерывном времени, как наиболее наглядное с точки зрения моделирования структуры НС и ее обучение в Simulink Matlab.

Пусть на интервале [0..T] на выходе НС наблюдается реализация случайного процесса с неизвестными параметрами $\omega_{ij}(t)$, требующими оценки.

Препятствием для формального обобщения рассмотренных выше дискретных алгоритмов на аналоговые, в которых используются функционалы от реализации случайных процессов, является то, что предела функции правдоподобия, когда размер выборки n неограниченно возрастает, не существует [5]. Однако для разработки аналоговых алгоритмов вместо несуществующего функционала правдоподобия можно ввести функционал отношения правдоподобия или его логарифм, который имеет вид

$$L[\ln \lim_{\substack{n\to\infty \\ T=const}} P_E(Y|X,W)/P_E(Y)] = L[\ln \Lambda(Y|X,W)],$$

откуда получаем оценку максимального правдоподобия величины W^* в точке $\omega_{ij}(t) = \omega^*_{ij}(t)$

$$\frac{\partial L[\Lambda(t)]}{\partial \omega_{ij}} = \int_0^T Z[E(\omega_{ij}(t))] \frac{\partial E(\omega_{ij}(t))}{\partial \omega_{ij}} dt = 0, \qquad (8)$$

где $Z[E] = -\frac{\partial \ln P_E(Y|X,W)}{\partial E}$.

Для робастных аналоговых алгоритмов вместо P_E и соответственно $Z[E]$ используем “наихудшие” в классах $\Re_n$ распределения. Например, для рассматриваемых классов $\Re_1, \Re_2$ аналоговые алгоритмы настройки НС (8) в точке оценки $\omega_{ij}(t) = \omega^*_{ij}(t)$ примут соответствующий вид:

$$\int_0^T \mathrm{sign}[E(\omega_{ij}(t))] \frac{\partial E(\omega_{ij}(t))}{\partial \omega_{ij}} dt = 0 \qquad (9)$$

и

$$\int_0^T E(\omega_{ij}) \frac{\partial E(\omega_{ij}(t))}{\partial \omega_{ij}} dt = 0\,, \tag{10}$$

где в формулах (8) – (10) обозначено $E(\omega_{ij}(t)) = G_j(y_j(t) - \sum_{i=0}^{q} \omega_{ij}(t) x_{ij}(t))$.

ЛИТЕРАТУРА

1. Уоссермэн Ф. Нейрокомпьютерная техника. – М.:Мир, 1992.
2. Аведьян В.Д. Алгоритмы настройки многослойных нейронных сетей // Автоматика и телемеханика, 1995, №4, С.106 – 118.
3. Фомин В.Н. Рекуррентное оценивание и адаптивная фильтрация. – М.: Наука. Гл. ред. физ.-мат. лит., 1984. – 288с.
4. Цыпкин Я.З. Основы информационной теории идентификации. – М.: Наука, 1984. – 320с.
5. Репин В.Г., Тартаковский Г.П. Статистический синтез при априорной неопределенности и адаптация информационных систем. – М.: Сов. радио, 1977. – 428с.