О СИНТЕЗЕ ФАКТОРОВ В ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЯХ

Н. А. Игнатьев
Национальный университет Узбекистана, Ташкент
e-mail: n_ignatev@rambler.ru

Algorithms of pre-processing of the various types of data from attribute spaces are considered for the purpose of minimization of the neural network configurations. These algorithms are applied to the problems of recognition "with a teacher".

Введение

Искусственные нейронные сети (НС) находят широкое применение там, где необходимо моделировать подобие человеческой интуиции и, как правило, трудно построить явные алгоритмы. Предобработка данных требуется для синтеза НС с минимальной конфигурацией и достижения высокой точности решения прикладных задач. Из теории и практики вычислений известно, что эти требования часто бывают диаметрально противоположными. Так, в задачах распознавания образов использование квадратичных решающих функций вместо линейных, с одной стороны, позволяет увеличить точность распознавания, а с другой — приводит к экспоненциальному росту объема вычислений.

Попытка увязать сложность конфигурации НС с наборами признаков, на которых эта конфигурация строится, предпринималась в [1]. Большое разнообразие условий, накладываемых на количественные признаки, включаемые в набор, отсутствие единого критерия отбора и трудности интерпретации процесса принятия решения в различных прикладных задачах сдерживали широкое практическое применение описанного в этой работе метода.

В настоящей статье рассматривается синтез факторов (комбинированных признаков) разнотипных признаковых пространств с целью построения нейронной сети с минимальной конфигурацией для корректных (не делающих ошибок) на обучающей выборке алгоритмов решения задач распознавания с учителем, дается формальное объяснение некоторых деталей процесса принятия решения. Вводятся новые способы определения:

— оценки межклассового различия и вклада признаков в разделение классов для отбора информативных наборов признаков;

— меры внутриклассового сходства (степени однородности) градаций номинальных признаков для вычисления синаптических весов нейронов и коррекции взвешенной суммы входных сигналов нейронов с учетом пропущенных значений (пропусков в данных) признаков.

Число комбинаций разнотипных признаков, используемых для синтеза НС с минимальной конфигурацией по алгоритмам из [2, 3], может определяться разными соображениями,

в том числе и экспертно-экспериментальными. Абсурдность разделения выборки на обучающую и контрольную для оценки качества распознавания НС, показанная в [2], служит веским аргументом для поиска новых критериев эффективности работы НС.

1. Синтез факторов разнотипного признакового пространства

Рассматривается стандартная задача распознавания образов с учителем. Задано множество объектов обучения $E_0 = \{S_1, ..., S_m\}$, содержащее представителей l непересекающихся классов $K_1, ..., K_l$ с описанием объектов в разнотипных признаковых пространствах. В описании объектов возможны пропуски данных.

Обозначим через I, J множество номеров соответственно количественных и номинальных признаков в описании допустимых объектов ($|I| + |J| = n$). Результат перемножения значений $k (k > 1)$ количественных признаков считается новым количественным признаком. Объединение номинальных признаков представляет номинальный признак с номером $p \notin J$, количество градаций которого ограничено сверху произведением числа градаций признаков, входящих в объединение, и значением $\min\limits_{1 \le j \le l} |K_j|$.

Считается, что синтез НС с минимальной конфигурацией аналогично [2] осуществляется в форме решения задачи о минимальном покрытии обучающей выборки E_0 объектами-эталонами множества $\Pi_j = \{S^1, ..., S^\alpha\}$, $\alpha \le m$, $\Pi_j \in E_0$, $j = 1, 2, ...$ Состав объектов покрытия Π_j зависит от порядка выбора объектов-кандидатов на удаление из E_0 процедурой "последовательное исключение". На множестве номинальных признаков вводится функция от трех переменных

$$f(r, a, b) = \begin{cases} \xi, & a = @ \text{ или } b = @, \\ 0, & a \ne b, \\ 1, & a = b, \end{cases}$$

где ξ — степень однородности градаций номинального признака в классе K_d и $S^r \in (K_d \cap \Pi_j)$; a, b — значения градаций; @ — код пропуска. Обозначим через I^*, J^* множества номеров исходных и комбинированных признаков со значениями соответственно в количественной и номинальной шкалах измерений. Положим, что объекты покрытия $\Pi_j \in E_0$ описываются признаками из $I^* \cup J^*$ и $|I^*| + |J^*| = \delta$.

Для распознавания принадлежности произвольно допустимого объекта $S = (b_1, ..., b_n)$ к классам $K_1, ..., K_l$ по Π_j производятся отображение $(b_1, ..., b_n) \to (y_1, ..., y_\delta)$ и вычисление

$$\varphi(S^r, S) = \sum_{i \in I^*} w_{ri} y_i + \sum_{i \in J^*} f(r, x_{ri}, y_i) w_{ri} + w_{r0}, \tag{1}$$

где $\{w_{r0}, w_{r1}, ..., w_{r\delta}\}$ — веса нейронов сети, определяемые по объекту-эталону $S^r = (x_{r1}, ..., x_{r\delta})$. Номер класса объекта S есть результат использования принципа "победитель забирает все" к значениям (1) на Π_j.

Будем считать, что $\{\eta_i\}_1^m$ — множество значений количественного признака $q \in I^*$ объектов из E_0, $A = (a_0, ..., a_l)$ — целочисленный вектор со значениями элементов: $a_0 = 0$, $a_l = m$, $a_r < a_{r+1}, r = \overline{1, l-1}$. Пусть

$$\eta_{i_1}, \ \eta_{i_2}, ..., \ \eta_{i_m} \tag{2}$$

— упорядоченная последовательность $\{\eta_i\}_1^m$, $\{u_1^1, ..., u_1^l, ..., u_l^1, ..., u_l^l\}$, набор целых чисел, элемент u_t^p в котором является количеством значений q-го признака объектов класса K_p в (2) с порядковыми номерами от $a_{t-1}+1$ до a_t.

Очевидно, что наилучшая разделимость классов, получаемая при переводе к номинальной шкале измерений, будет тогда, когда значения номинального признака одинаковы внутри каждого класса и не совпадают ни с одним значением из других классов, а число градаций признака равно числу классов.

Все значения количественного признака $q \in I^*$ в (2) с номерами от a_{t-1} до a_t, $t = \overline{1,l}$ согласно критерию

$$\frac{\left(\sum\limits_{p=1}^{l}\sum\limits_{i=1}^{l}(u_i^p-1)u_i^p\right)\left(\sum\limits_{i=1}^{l}|K_i|(m-|K_i|)\right)}{\left(\sum\limits_{p=1}^{l}\sum\limits_{i=1}^{l}u_i^p(m-|K_i|-\sum\limits_{j=1}^{l}u_j^p+u_i^p)\right)\left(\sum\limits_{i=1}^{l}|K_i|(|K_i|-1)\right)} \to \min_{\{A\}} \tag{3}$$

считаются эквивалентными в номинальной шкале измерений.

Обозначим через p число градаций признака $c \in J^*$, g_{dc}^t, $\overline{g_{dc}^t}$ — количество значений t-й $(1 \le t \le p)$ градации c-го признака в описании объектов соответственно класса K_d и его дополнения CK_d, $\theta_{dc}, \overline{\theta_{dc}}$ — число значений c-го признака без пропусков соответственно в K_d и CK_d, $l_{dc}, \overline{l_{dc}}$ — число градаций c-го признака соответственно в K_d и CK_d. Межклассовое различие по c-му признаку определяется как величина

$$\lambda_c = 1 - \frac{\sum\limits_{i=1}^{l}\sum\limits_{t=1}^{p} g_{ic}^t \overline{g_{ic}^t}}{\sum\limits_{i=1}^{l}(\theta_{ic}-l_{ic}+1)(\overline{\theta_{ic}}-\overline{l_{ic}}+1)+(\min(l_{ic},\overline{l_{ic}})-1)}. \tag{4}$$

Степень однородности (мера внутриклассового сходства) значений градаций c-го признака по классу K_d вычисляется по формуле

$$\beta_{dc} = \frac{\sum\limits_{t=1}^{l_{dc}} g_{dc}^t(g_{dc}^t-1)}{(\theta_{dc}-l_{dc}+1)(\theta_{dc}-l_{dc})} \tag{5}$$

и используется в качестве значения ξ функции $f(r,a,b)$ в (1). С помощью (4),(5) стало возможным определять "индивидуальные" веса номинального признака в разных классах. Так, для объекта $S^r \in \Pi_j \cap K_d$ вес c-го признака в (1) вычисляется по формуле $w_{rc} = \lambda_c \beta_{dc}$.

Другим применением значения (4) является использование его в качестве показателя для сравнения при отборе информативных комбинаций разнотипных признаков. Множество сравниваемых комбинаций признаков может быть получено с помощью переборных или генетических алгоритмов. В качестве побочного эффекта от объединения номинальных признаков отметим следующее: возрастает вероятность того, что комбинированный признак произвольного допустимого объекта S содержит градации, отсутствующие у объектов обучения. Преобразование количественных признаков по критерию (3) позволяет синтезировать новые номинальные признаки как комбинации из количественных и номинальных признаков.

Процесс синтеза нового количественного признака x_q, $q \in I^*$, в общем виде представляется как

$$x_q = \psi_1(x_{i_1}) \times \psi_2(x_{i_2}) \times ... \times \psi_k(x_{i_k}),$$

где $\psi_d(x_{i_t})$ — преобразование (в том числе и тождественное) признака x_{i_t} в определенную количественную шкалу измерений. Примером преобразования, меняющего порядок следования (2) на обратный, служит уравнение

$$\psi(x) = \frac{x_{\max} - x}{x_{\max} - x_{\min}}, \tag{6}$$

в котором $x_{\max}, x_{\min}$ — соответственно максимальное и минимальное значения признака x. Исследование и обоснование выбора различных преобразований при синтезе количественных признаков в данной работе не рассматриваются.

Аналогично [1] выбор весов количественных признаков в (1) осуществляется с помощью взвешенной евклидовой метрики

$$\rho(S, S_i) = \sqrt{\sum_{c \in I^*} v_c^2 (y_c - x_{ic})^2}, \tag{7}$$

где $S = (y_1, ..., y_\delta),\ S_i = (x_{i1}, ..., x_{i\delta})$. С этой целью для каждого признака $x_c, c \in I^*$, по критерию (3) определяются значения градаций в номинальной шкале измерений и вычисляется

$$v_c = \frac{\sqrt{\lambda_c \dfrac{\sum\limits_{d=1}^{l} \sum\limits_{t=1}^{l_{dc}} g_{dc}^t (g_{dc}^t - 1)}{\sum\limits_{d=1}^{l} (\theta_{dc} - l_{dc} + 1)(\theta_{dc} - l_{dc})}}}{\max\limits_{1 \le j \le m} x_{jc} - \min\limits_{1 \le j \le m} x_{jc}}.$$

Значения весов количественных признаков объектов покрытия $\Pi_j = \{S^1, ..., S^\alpha\}$ в (1) определяются как $w_{ri} = v_i^2 x_{ri}$ и $w_{r0} = -\sum\limits_{i \in I^*} w_{ri}^2/2$.

2. Критерии оценки качества синтеза признаков и вычислительный эксперимент

При выборе критериев оценки качества синтеза признаков имеет смысл отдельно рассматривать случаи, когда признаковое пространство представлено: а) количественными признаками; б) разнотипными признаками.

В первом случае на множестве объектов обучения E_0 определяется линейная оболочка $L(E_0)$ [1], являющаяся подмножеством граничных объектов классов по метрике (7). Различные наборы признаков сравниваются по критерию

$$\frac{\sum\limits_{S_i \in L(E_0)} \rho(S_i, S_i^*)}{\delta |L(E_0)|} \to \max_{E_0},$$

в котором $S_i^* \in CK_d,\ d = \overline{1,l}$, — ближайший (по метрике (7)) объект к $S_i \in K_d$, δ — число признаков в наборе. Предпочтительным считается тот набор признаков, на котором получено максимальное в среднем расстояние между линейными оболочками классов.

При наличии пропусков в данных для анализа комбинаций количественных признаков целесообразно использовать преобразование по критерию (3) и определять значение

вклада каждого признака $p \in I^*$ в разделение классов как

$$\lambda_p = \frac{\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{u_p} z_{pj}^i(z_{pj}^i - 1)}{\sum\limits_{i=1}^{l} b_{ip}(b_{ip}-1)} - \frac{\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{u_p} z_{pj}^i \overline{z_{pj}^i}}{\sum\limits_{i=1}^{l} b_{ip}\overline{b_{ip}}}, \tag{8}$$

где z_{pj}^i, $\overline{z_{pj}^i}$ — количество значений j-й градаций p-го признака соответственно класса K_i и его дополнения $CK_i = E_0 \backslash K_i$; u_p — число градаций p-го признака; b_{ip}, $\overline{b_{ip}}$ — число значений p-го признака без пропусков соответственно в K_i и CK_i. Упорядочение множества значений $\{\lambda_p\}$ позволяет производить направленный отбор информативных наборов признаков. Для заполнения пропусков значений количественных признаков можно использовать хорошо известные и описанные в научной литературе методы. Для анализа качества заполнения тем или иным методом рекомендуется сравнивать значения (8), полученные до и после заполнения пропусков.

Для разнотипных признаков (второй случай) критерием качества служит оценка сложности решающей функции на локально-оптимальном покрытии обучающей выборки объектами-эталонами, используемая в [1]. Оценка сложности вычисляется как произведение числа объектов-эталонов локально-оптимального покрытия обучающей выборки на размерность признакового пространства и базируется на таком фундаментальном понятии, как емкость класса решающих функций в методе структурной минимизации риска [4].

Значения (4),(5) могут быть использованы для интерпретации экспериментальных табличных данных в терминах нечетких логик. Всегда нужно помнить, что эксперта-исследователя чаще всего интересует не только результат распознавания, но и объяснение того, как этот результат получился.

Обозначим через $R^k(t)$ пространство из t признаков, в котором $k(k \geq 1)$ определяет максимальное число исходных признаков, используемое для синтеза комбинированного признака. Для вычислительного эксперимента были взяты медицинские данные из [5], содержащие описания 177 объектов с помощью 29 количественных признаков. Объекты выборки разделены на два непересекающихся класса: класс 1 — контрольная группа (111 человек), класс 2 — больные гипертонией (66 человек). Количество пропусков в данных равно 7.23 %, и для заполнения их при выборе минимальной конфигурации НС использовались средние значения признаков в классах.

Перечень из 29 признаков, упорядоченный по мере уменьшения их вклада (8) в разделение объектов классов, выглядел следующим образом:

1) среднее артериальное давление;
2) систолическое артериальное давление;
3) диастолическое артериальное давление;
4) пульсовое артериальное давление;
5) размер полости левого предсердия;
6) возраст;
7) конечный систолический размер левого желудочка;
8) конечный систолический объем левого желудочка;
9) удельное периферическое сопротивление;
10) конечный диастолический объем левого желудочка;
11) конечный диастолический размер левого желудочка;
12) вес;

13) индекс Кердо;
14) фракция выброса;
15) степень укорочения переднезаднего размера левого желудочка в систолу;
16) ударный объем;
17) рост;
18) коэффициент К1;
19) минутный объем;
20) систолический показатель;
21) длительность интервала QT на ЭКГ;
22) длительность систолы;
23) длительность диастолы;
24) коэффициент К2;
25) длительность интервала QRS на ЭКГ;
26) частота пульса;
27) сердечный индекс;
28) длительность интервала PQ на ЭКГ;
29) длительность интервала RR на ЭКГ.

Для сравнительного анализа были рассмотрены два пространства: $R^1(29)$ из 29 исходных признаков и $R^2(7)$, содержащие семь парных комбинаций признаков, вклад (8) каждой из которых больше, чем у среднего артериального давления. Синтез признаков для $R^2(7)$ осуществлялся с помощью тождественного $\psi_{\text{тож}}(x) = x$ и обратного $\psi_{\text{обр}}(*)$ преобразований (6). Перечень из семи комбинированных признаков, расположенных в порядке убывания значений (8), был получен как результат произведения следующих преобразований исходных признаков:

— $\psi_{\text{тож}}$ (рост) $\times$ $\psi_{\text{обр}}$ (систолическое артериальное давление);

— $\psi_{\text{тож}}$ (диастолическое артериальное давление) $\times$ $\psi_{\text{обр}}$ (среднее артериальное давление);

— $\psi_{\text{тож}}$ (рост) $\times$ $\psi_{\text{обр}}$ (среднее артериальное давление);

— $\psi_{\text{тож}}$ (систолическое артериальное давление) $\times$ $\psi_{\text{тож}}$ (размер полости левого предсердия);

— $\psi_{\text{тож}}$ (размер полости левого предсердия) $\times$ $\psi_{\text{тож}}$ (среднее артериальное давление);

— $\psi_{\text{тож}}$ (сердечный индекс) $\times$ $\psi_{\text{тож}}$ (удельное периферическое сопротивление);

— $\psi_{\text{тож}}$ (систолическое артериальное давление) $\times$ $\psi_{\text{тож}}$ (среднее артериальное давление).

Эффект от предобработки данных в виде количества объектов покрытия при синтезе НС с минимальной конфигурацией приведен в таблице. В пространстве с евклидовой метрикой для каждого объекта покрытия $S^r \in \Pi_j$, $S^r = (x_{r1}, ..., x_{rt})$, веса в (1) вычислялись по формулам $w_{ri} = x_{ri}$, $w_{r0} = -\sum_{i=1}^{t} w_{ri}/2$. Для выбора локально-оптимальных покрытий в $R^1(29)$ и $R^2(7)$ процедурой "последовательное исключение" использовался один и тот же порядок подачи объектов-кандидатов на удаление с номерами от 1 до 177. Трудоемкость вычисления информативных наборов признаков выразилась в форме линейной зависимо-

Пространство	$R^1(29)$	$R^2(7)$
С евклидовой метрикой	22	13
С метрикой (7)	18	9

сти между затратами процессорного времени и числом признаков, предъявляемых для отбора.

Синтез НС в обобщенном признаковом пространстве, определяемом информативным набором комбинированных признаков, позволил в несколько раз снизить сложность (произведение числа эталонов покрытия на размерность признакового пространства) решающих функций по сравнению с аналогичными показателями для исходного признакового пространства.

Список литературы

[1] Игнатьев Н.А. Выбор минимальной конфигурации нейронных сетей // Вычисл. технологии. 2001. Т. 6, № 1. С. 23–28.

[2] Игнатьев Н.А. Извлечение явных знаний из разнотипных данных с помощью нейронных сетей // Вычисл. технологии. 2003. Т. 8, № 2. С. 69–73.

[3] Игнатьев Н.А., Мадрахимов Ш.Ф. О некоторых способах повышения прозрачности нейронных сетей // Вычисл. технологии. 2003. Т. 8, № 6. С. 31–37.

[4] Прикладная статистика: Классификация и снижение размерности: Справочное издание / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. М.: Финансы и статистика, 1989.

[5] Ignat'ev N.A., Adilova F.T., Matlatipov G.R., Chernysh P.P. Knowledge discovering from clinical data based on classification tasks solving // MediNFO. Amsterdam: IOS Press, 2001. P. 1354–1358.

Поступила в редакцию 28 октября 2003 г.,
в переработанном виде — 24 декабря 2004 г.