УДК 004.[023+622+89]

© 2009 г. **В.В. Аюев**, канд. техн. наук,
П.А. Карпухин
(Калужский филиал Московского государственного технического университета им. Н.Э.Баумана)

КЛАСТЕРНЫЙ МЕТОД ПОДБОРА ПАРАМЕТРОВ И ОБУЧЕНИЯ НА НЕПОЛНЫХ ДАННЫХ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ХЕХТ-НИЛЬСОНА

Описывается применение алгоритма плотностной статической кластеризации к решению задачи эффективного обучения нейронной сети Хехт-Нильсона. Использованы оригинальные механизмы очистки выборки от «шума», восстановления поврежденных данных, бисекционного выборочного поиска кластеров с последующей установкой в них центров нейронов Кохонена.
Ключевые слова: плотностная кластеризация, восстановление пропусков данных, нейронная сеть Хехт-Нильсона, бисекционная выборочная кластеризация.

Введение

Локально-аппроксимирующие искусственные нейронные сети (ИНС), традиционно рассматривавшиеся в качестве связующего звена между многослойными персептронами и РБФ-сетями [1], находят все более широкое практическое применение ввиду как увеличения размерности решаемых задач, так и объема используемых обучающих множеств.

Одна из наиболее известных подобных нейроархитектур – ИНС Хехт-Нильсона [2], основанная на кластеризации входного пространства и соотнесении каждому кластеру некоторой локальной аппроксимации, – обладает рядом преимуществ по сравнению с широко распространенными многослойными персептронами: высокой скоростью обучения, отсутствием эффекта переобучения и простой структурой. Сети Хехт-Нильсона (СХН) также несвойственна проблема плохой обусловленности интерполяционных матриц, характерная для РБФ-сетей [3]. Вместе с тем СХН имеет ряд ограничений, определяемых ее архитектурой, среди которых следует выделить длительность процесса кластеризации и высокую чувствительность к выбросам. Модификация СХН на основе внедрения высокоразмерных карт Кохонена и разработанного алгоритма перекластеризации, осуществленная в работе [4], позволила значительно сократить время обучения, однако принципиально не решила обозначенные проблемы.

Другой достаточно общей проблемой, ограничивающей применение нейро-

сетевых парадигм, являются пропуски в исходных данных, образующих обучающую выборку. Здесь и далее под информационными пропусками понимается отсутствие сведений о значении некоторого элемента вектора в выборке, за исключением его типа, определяемого типом атрибута. В случае, когда пропуски имеют регулярный характер относительно некоторого подмножества атрибутов, проблема может быть решена удалением этих атрибутов из всех векторов обучающего множества [5]. При случайном распределении пропусков как по атрибутам, так и по векторам могут быть применены различные подходы на основе деревьев решений [6], автоассоциативных ИНС [7], вычисления ближайших общих соседей [8], а также линейно-регрессионных методов [9]. Перечисленным подходам в различной степени присущи следующие недостатки, описанные в работах Фуджикавы и др. [10], Ву и др. [11]: высокая вычислительная сложность, спецификация по типу атрибутов, необходимость в накоплении БД с фоновыми знаниями, а также трудности с внедрением в нейросетевую модель.

Принимая во внимание проблемы описанных выше подходов, в настоящем исследовании предлагается использовать алгоритм статической кластеризации, введенный в работе [12], с целью компенсации информационной неполноты обучающей выборки, а также для удаления «информационного шума», выявляемого в процессе кластеризации. Последующая обработка полученных кластеров выборочным бисекционным алгоритмом *K*-средних [13] позволяет выявить заданное архитектурой ИНС количество кластеров и установить в них центры нейронов Кохонена без осуществления длительной процедуры самоорганизации. Структурная схема предлагаемого подхода приведена на рис. 1.

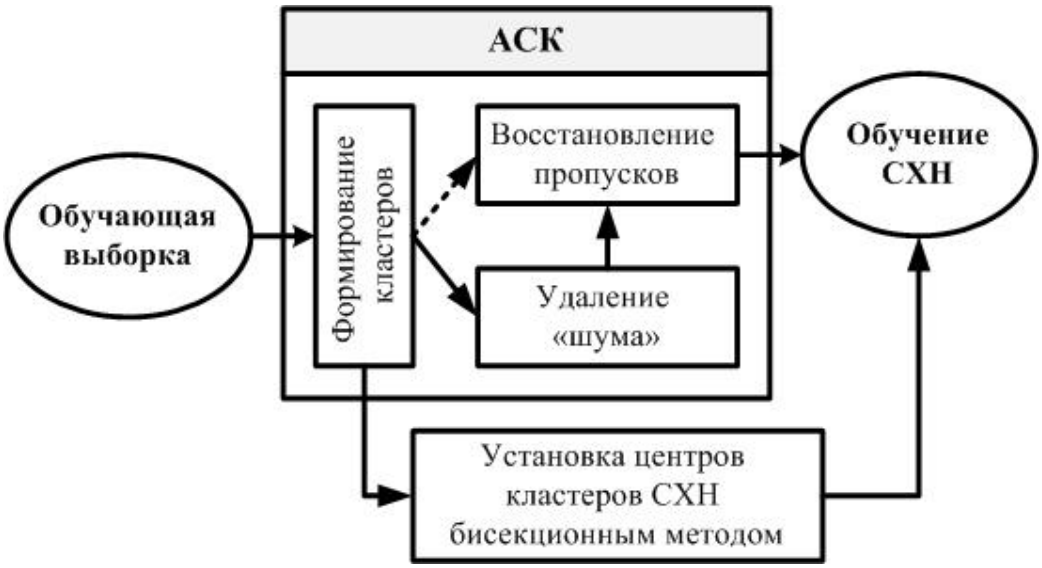

Рис. 1. Структурная схема восстановления данных и установки центров нейронов Кохонена при обучении СХН выбранной архитектуры.

Нейросетевая модель Хехт-Нильсона

Искусственная нейронная сеть Хехт-Нильсона разбивает *N*-мерную область входных сигналов на *L* кластеров и ставит в соответствие каждому кластеру выходной вектор размерности *M*, соответствующий значению аппроксимируемой

функции (рис. 2). Промежуточный слой сети состоит из $N_{1,l}$ (l = 1, 2, ..., L) нейронов Кохонена [14], каждый из которых связан произведением с нелинейным нейроном f_l. В свою очередь все нейроны промежуточного слоя связаны с линейными нейронами выходного слоя $N_{2,m}$ (m = 1, 2, ..., M).

Значения выходных сигналов нелинейных нейронов промежуточного слоя описываются выражением:

$$y_{1,l}^{(f)} = y_{1,l} \cdot f_l\left(\sum_{n=1}^{N} x_n w_{0,n;1,l}^{(f)} \right),\ l = 1,2,\ldots,L, \tag{1}$$

где f – верхний индекс соответствует элементам, относящимся к нелинейным нейронам. Значения нейронов выходного слоя определяются как:

$$y_m = \sum_{l=1}^{L} y_{1,l}\left[w_{1,l;2,m} + w_{1,l;2,m}^{(f)} \cdot f_l\left(\sum_{n=1}^{N} x_n w_{0,n;1,l}^{(f)} \right) \right],\ m = 1,2,\ldots,M. \tag{2}$$

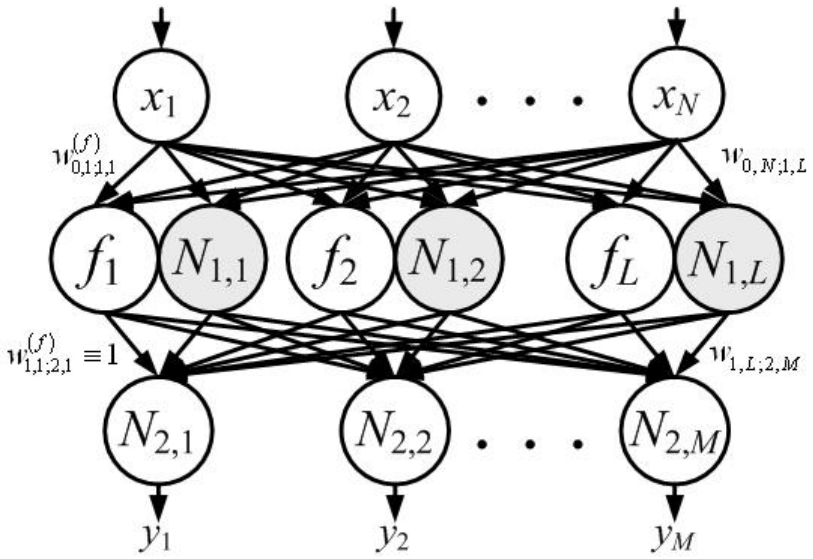

Рис. 2. Структура сети Хехт-Нильсона.

Пусть в рассматриваемой СХН функции поощрения гомогенны и имеют дискретный вид, т.е. выходные значения нейронов Кохонена отличны от нуля лишь для нейрона-победителя с индексом ϕ:

$$y_{1,l} = \begin{cases} 1, & l = \tau; \\ 0, & l \neq \tau \end{cases}, \tag{3}$$

а веса связей между нелинейными нейронами промежуточного слоя и нейронами выходного слоя равны 1:

$$w_{1,l;2,m}^{(f)} = 1,\ \ l = 1,2,\ldots,L,\ \ m = 1,2,\ldots,M. \tag{4}$$

Из допущений (3), (4) следует, что выражение (2), описывающее работу ИНС, примет вид:

$$y_m = w_{1,\tau;2,m} + f_l\left(\sum_{n=1}^{N} x_n w_{0,n;1,\tau}^{(f)} \right),\quad m = 1,2,\ldots,M. \tag{5}$$

Обучение СХН проводится в два этапа. Первый этап, традиционно реализуемый посредством самоорганизации нейронов Кохонена, связан с нахождением

весов $w_{0,h;1,i}$, представляющих центр соответствующего кластера. В нашей работе процесс самоорганизации заменен двухуровневой кластеризацией, алгоритм которой приведен в разделе 2: сначала осуществляется поиск больших пространственных кластеров, затем кластеры разбиваются бисекционным алгоритмом на заданное количество (L) меньших кластеров, центры которых соответствуют искомым весовым коэффициентам.

Второй этап состоит в нахождении весовых коэффициентов $w^{(f)}_{0,n;1,\phi}$ и $w_{1,\phi;2,m}$ в соответствии с парадигмой обратного распространения ошибки [15]:

$$w_{1,\tau;2,m}(t+1) = w_{1,\tau;2,m}(t) + \eta(d_m - y_m),$$
$$w^{(f)}_{0,n;1,\tau}(t+1) = w^{(f)}_{0,n;1,\tau}(t) + \eta x_n f' \sum_{m=1}^{M}(d_m - y_m), \quad 0 < \eta \le 1, \tag{6}$$

где d_m – целевой m-й выход СХН; $з$ – коэффициент скорости обучения; f'– производная активационной функции по активационному потенциалу ϕ-го нелинейного нейрона, связанного с победившим нейроном Кохонена $N_{1,\phi}$.

Формирование обучающего множества по неполным исходным данным

Решение проблемы восстановления информационных пропусков в обучающей выборке осуществлено на основе алгоритма статической кластеризации (АСК), исследованного в ряде работ [12, 16 – 18].

В основе АСК лежит плотностной подход к кластеризации DBSCAN, впервые описанный в работе [19], с выделением трех типов объектов: внутренних, граничных и шума. Разработанный для кластеризации больших баз данных, DBSCAN превосходит по соотношению скорости работы по результирующему качеству кластеризации в высокоразмерных пространствах большую часть известных алгоритмов [20]. Ключевым фактором, повлиявшим на выбор подобного решения, является его способность эффективно выявлять кластеры различных форм, размеров и плотности [21].

Следует отметить, что хотя использованный подход и уступает по скорости работы некоторым более современным методам, – например, DENCLUE [22] и CUBN [23], он требует установки значений меньшего числа параметров; в отличие от методов на основе минимальных покрывающих деревьев результат не зависит от выбора начального объекта кластеризации; в отличие от гибридных иерархических методов, – например, CURE [24], относительная плотность расположения векторов в кластере остается постоянной, что является принципиально важным свойством ввиду последующей обработки бисекционным выборочным алгоритмом.

Мера расстояния. Важнейшим вопросом кластеризации данных является корректный выбор меры для вычисления расстояния между объектами кластеризации. Учитывая произвольную природу данных, составляющих обучающее множество **X**, включающее P векторов, а также отсутствие ограничения на их размерность, вместо классической меры Минковского и ее частных случаев, применение которых представляется нецелесообразным в ряде исследований [25, 26], в качестве меры расстояния между элементами множества предлагается исполь-

зовать выражение вида:

$$dst(\mathbf{x}_i, \mathbf{x}_j) = \sum_{n=1}^{N} \alpha^{(n)} d_{i,j}^{(n)} \delta_{i,j}^{(n)} \Big/ \sum_{n=1}^{N} \delta_{i,j}^{(n)}, \quad (7)$$

где

$$d_{i,j}^{(n)} = \begin{cases} 0, & x_{i,n} = x_{j,n}, \\ 1, & x_{i,n} \neq x_{j,n}, \end{cases} \quad (8)$$

для категорных атрибутов, или

$$d_{i,j}^{(n)} = \frac{\left| x_{i,n} - x_{j,n} \right|}{\max\limits_{p=1\ldots P_n} x_{p,n} - \min\limits_{p=1\ldots P_n} x_{p,n}}, \quad (9)$$

для непрерывных атрибутов. Здесь $\mathbf{x}_i$ – i-й вектор обучающего множества; $б^{(n)}$ – фактор влияния n-го атрибута обучающего множества; $d^{(n)}_{i,j}$ – расстояние между n-ми атрибутами векторов i и j; $\partial^{(n)}_{i,j}$=0, если имеется пропуск в n-м атрибуте для одного из сравниваемых векторов i или j; $x_{i,n}$ – величина n-го атрибута i-го вектора обучающего множества; P_j – количество векторов обучающего множества, содержащих j-й атрибут без пропуска информации.

Особенностью данного решения является возможность непосредственной работы с данными, содержащими пропуски информации, без необходимости предварительной оценки их значений точным, интервальным или вероятностным способами [27].

ACK. Алгоритм статической кластеризации использует информацию о ближайших общих соседях (БОС), наряду с данными о взаимном расположении векторов в пространстве для оценки степени схожести векторов друг с другом. Наиболее похожие группы векторов образуют ядра кластеров – множество внутренних векторов. К ядрам на основе БОС-близости или расстояния добавляются векторы, не обладающие достаточным показателем БОС – так называемые граничные векторы. Наконец, векторы, имеющие самые низкие показатели БОС, т.е. наименьшую степень схожести со всеми остальными векторами в выборке, считаются «шумом» – паразитными объектами, подлежащими удалению из выборки.

Формально алгоритм состоит из следующих шагов:

1. Выбор величин K – числа анализируемых ближайших соседей для каждого вектора; R – минимального числа ближайших общих соседей у двух векторов, достаточного для того, чтобы считать их принадлежащими некоторому кластеру; R^* ($K > R^* > R$) – минимального числа общих соседей у двух векторов, достаточного для того, чтобы считать их принадлежащими одному кластеру.

2. Объединение векторов в кластеры по принципу перекрестного соседства: два вектора входят в один кластер, если между ними можно выстроить цепь соседей, каждый из которых будет иметь показатель *SNN*, больший или равный R^*.

3. Добавление к созданным кластерам векторов, имеющих показатель *SNN* ,больший или равный R, на основе принципа максимизации показателя соседства: вектор добавляется к кластеру, с объектом которого у него максимальная величина *SNN*. В случае, если вектор имеет равный показатель соседства с векторами, из

двух и большего количества кластеров выбирается такой кластер, объект которого находится ближе к данному вектору в смысле расстояния (7)-(9).

Замена информационных пропусков. Результатом работы алгоритма АСК является формирование S кластеров, содержащих большую часть (в зависимости от величины R) элементов выборки. Для устранения пропусков в этих векторах предлагается использовать методы групповой и единичной замен в зависимости от типов атрибутов, содержащих пропуск информации:

$$x_{i,j}^{ms,(C)} = \left[\sum_{h=1}^{|\mathbf{C}_s|} SNN\left(\mathbf{x}_i^{(C)}, \mathbf{x}_h^{(C)}, K\right) x_{i,j}^{(C)} \right] \Bigg/ \sum_{h=1}^{|\mathbf{C}_s|} SNN\left(\mathbf{x}_i^{(C)}, \mathbf{x}_h^{(C)}, K\right), \tag{11}$$

для данных, описываемых непрерывными атрибутами, и

$$x_{i,j}^{ms,(C)} = x_{p,j}^{(C)} : \begin{cases} \min\limits_{p=1\ldots P_j^{(C)}} \left\{ \dfrac{dst\left(\mathbf{x}_i^{(C)}, \mathbf{x}_p^{(C)}\right)}{SNN\left(\mathbf{x}_i^{(C)}, \mathbf{x}_p^{(C)}, K\right)} \right\}, & dst\left(\mathbf{x}_i^{(C)}, \mathbf{x}_p^{(C)}\right) \neq 0, \\ \max\limits_{p=1\ldots P_j^{(C)}} SNN\left(\mathbf{x}_i^{(C)}, \mathbf{x}_p^{(C)}, K\right), & dst\left(\mathbf{x}_i^{(C)}, \mathbf{x}_p^{(C)}\right) = 0, \end{cases} \tag{12}$$

для данных, описываемых категорными атрибутами. Здесь $x^{ms,(C)}{}_{i,j}$ – замещаемый (содержащий пропуск) j-й атрибут i-го вектора $\mathbf{x}^{(C)}{}_i$ из обучающей выборки, который принадлежит кластеру $\mathbf{C}_s$; $SNN(\mathbf{x}_i, \mathbf{x}_j, K)$ – количество ближайших общих соседей среди первых K из P-2 возможных соседей между векторами i и j; $x^{(C)}{}_{h,j}$ – j-й (без пропусков) атрибут h-го вектора $\mathbf{x}^{C}{}_i$ из обучающей выборки, который принадлежит кластеру $\mathbf{C}_s$; $P^{(C)}{}_j$ – количество векторов в кластере $\mathbf{C}_s$ с j-м атрибутом без пропусков информации; $S \geq s \geq 1$.

В случае, если обрабатываемый кластер содержит пропуски информации во всех векторах для некоторого атрибута, выражение (12) применяется к ближайшему элементу из другого кластера, не содержащему пропуск в соответствующем атрибуте.

Формирование замен (11), (12) производится в соответствие с методами, экспериментально показавшими лучшее соотношение скорости и качества (в смысле результирующей ошибки) раздельно для непрерывных и категорных атрибутов. Более подробно с результатами сравнительного анализа различных подходов к замене пропусков данных можно ознакомиться в работе [16].

Установка центров нейронов Кохонена в СХН

Целью процедуры кластеризации в СХН является установка весовых коэффициентов нейронов Кохонена, которым соответствуют центры глобулярных кластеров.

Кластеры, сформированные в результате применения алгоритма АСК, во-первых, обладают сложной пространственной формой, следовательно, координаты их центров непригодны для непосредственной инициализации весовых коэффициентов нейронов Кохонена, а во-вторых, содержат большое количество векторов. Предлагаемое в настоящей работе решение состоит в формировании селективных подмножеств для каждого кластера, последующем разбиении полученных

подмножеств бисекционным методом, где выборочным и терминальным условием выступает расстояние между объектами, вычисляемое в соответствии с (7) – (9), и консолидации полученных результатов. Бисекционный алгоритм, формирующий глобулярные кластеры, был выбран в качестве основы ввиду простоты и высокой скорости работы [21].

Селективность подмножеств позволяет сократить количество обрабатываемых векторов в кластере посредством выборки лишь небольшого числа «представителей» этого кластера. Подобное решение становится возможным ввиду одинаковой плотности расположения векторов в кластере – фактического равномерного распределения, – которая допускает случайный выбор элементов в селективное подмножество. Для устранения эффекта неудачного выбора элементов, а также их малого количества предлагается использовать несколько селективных подмножеств с целью выбора наилучшего разделения кластера надвое.

Бисекционный выборочный алгоритм. Предлагаемый алгоритм, описанный далее, обрабатывает классический случай, когда исходное количество кластеров S меньше необходимого L, установленного архитектурой СХН.

1. Выбор количества селективных подмножеств (A) и их величины (B).
2. Выбор из S кластеров наибольшего ($\mathbf{C}_{max}$) для последующего разбиения:

$$\mathbf{C}_{max} = \max_{p=1\ldots S} \frac{1}{|\mathbf{C}_s|} \sum_{h=1}^{|\mathbf{C}_s|} dst\left(\mathbf{м}_s, \mathbf{c}_h^{(s)}\right), \tag{13}$$

$$\boldsymbol{\mu}_s = \frac{1}{|\mathbf{C}_s|} \sum_{h=1}^{|\mathbf{C}_s|} \mathbf{c}_h^{(s)}. \tag{14}$$

Здесь $\boldsymbol{\mu}_s$ – центроид кластера $\mathbf{C}_s$; $\mathbf{c}_h^{(s)}$ – h-й вектор, принадлежащий кластеру $\mathbf{C}_s$.

3. Формирование A подмножеств $\mathbf{C}^1{}_{max}$, $\mathbf{C}^2{}_{max}$, ..., $\mathbf{C}^A{}_{max}$ случайным выбором B элементов из $\mathbf{C}_{max}$ и инициализация центроидов будущих пар кластеров $\mathbf{M} = \{(\boldsymbol{\mu}^{a,1}{}_{max}, \boldsymbol{\mu}^{a,2}{}_{max})\}$, где $a = 1, 2, ..., A$. Соотнесение каждого вектора в образованных подмножествах одному из кластеров $\mathbf{C}^{a,1}{}_{max}$ или $\mathbf{C}^{a,2}{}_{max}$ на основе их близости в смысле (7) – (9) к центроидам с последующим пересчетом $\mathbf{M}$ в соответствии с (14).

4. Повторение п. 3 до тех пор, пока координаты центроидов не перестанут изменяться:

$$\sum_{s=1}^{S} dst\left(\mathbf{м}_s(t-1), \mathbf{м}_s(t)\right) < \varepsilon, \tag{15}$$

где $e \geq 0$ – максимально допустимая величина взвешенного отклонения центроид на двух последовательных итерациях; $\boldsymbol{\mu}_s(t-1)$ – координаты s-го центроида, полученного на предыдущей итерации; $\boldsymbol{\mu}_s(t)$ – координаты s-го центроида, полученного на текущей итерации.

5. Выбор результирующих координат центроидов двух кластеров $\mathbf{C}^1{}_{max}$ и $\mathbf{C}^2{}_{max}$, образуемых из кластера $\mathbf{C}_{max}$ таким образом, чтобы расстояние от центра кластера до всех принадлежащих ему объектов было минимальным.

6. Сохранение объектов из $\mathbf{C}_{max}$ в двух новых кластерах и увеличение счетчика S.

7. Если $S < L$, переход к п.2, иначе установить координаты нейронов Кохонена в центроиды соответствующих кластеров, рассчитанные согласно (14).

Заметим, что в СХН отсутствует требование о топологически непрерывном упорядочивании весовых коэффициентов, – следовательно, соотнесение центров кластеров нейронам Кохонена (п.7) может проводиться в произвольном порядке.

Слияние кластеров. Установка большого значения параметра R^* и малого R при относительно большом K ($K >> R$) может приводить к формированию АСК значительного числа кластеров небольшого размера с ядром из нескольких векторов. В этом случае применяется метод кластерного слияния на основе простой минимизации совокупного расстояния до центроида у объединяемых кластеров: выбирается пара кластеров, объединение которых приведет к формированию кластера с минимальной суммой расстояний от векторов до его центра:

$$\mathbf{C}_i = \mathbf{C}_i \cup \mathbf{C}_j : \quad \min_{\substack{i,j=1,\dots S \\ i \neq j}} \frac{1}{\left|\mathbf{C}_i\right| + \left|\mathbf{C}_j\right|} \left(\sum_{g=1}^{\left|\mathbf{C}_i\right|} dst\left(\text{ŕĶ}, \mathbf{c}_g\right) + \sum_{h=1}^{\left|\mathbf{C}_i\right|} dst\left(\text{ŕĶ}, \mathbf{c}_h\right) \right), \tag{16}$$

где $\boldsymbol{\mu}_*$ – центроид кластера, образуемого в результате объединения $\mathbf{C}_i$ и $\mathbf{C}_j$, рассчитываемый в соответствие с выражением:

$$\boldsymbol{\mu}_* = \frac{1}{\left|\mathbf{C}_i\right| + \left|\mathbf{C}_j\right|} \left(\sum_{g=1}^{\left|\mathbf{C}_i\right|} \mathbf{c}_g + \sum_{h=1}^{\left|\mathbf{C}_i\right|} \mathbf{c}_h \right). \tag{17}$$

Экспериментальный анализ

Анализ разработанной системы был осуществлен на примере нейросетевого решения задачи классификации, исходные данные которой, полученные в ходе классификации биологических видов галиотис и измерения их физиологических параметров, представлены в открытом репозитории UCI [28]. В качестве критериев эффективности выступали временные и качественные – среднеквадратичное отклонение (СКО) целевых значений от действительных – показатели работы системы. Результаты сравнивались с аналогичными показателями, полученными для обобщенных сетей радиально-базисных функций [15] стандартной архитектуры с 25 базисными нейронами (далее РБФ) и трехслойных персептронов (8-20-10-1) с сигмоидальными активационными функциями и линейным выходным нейроном (далее МП). Для работы АСК были выбраны параметры $R = 30$, $K = 50$. Нейроархитектура СХН с внедрённым АСК, описанная в настоящей работе, имела 25 пар элементов в скрытом слое (далее КСХН).

Временные оценки, приводимые далее, получены на аппаратной платформе Intel Pentium QX9550 с 8 Гб ОЗУ под управлением ОС Windows.

Формирование выборок. Экспериментальная выборка основана на открытой статистической базе «Abalone», широко применяемой для исследования эффективности вычислительных моделей, обрабатывающих разнородные по типу данные [http://archive.ics.uci.edu/ml/datasets/Abalone]. Исходная выборка включала 4177 векторов, содержащих 9 атрибутов категорного и непрерывного типов, описывающих пол, длину, рост, диаметр, полный вес, вес раковины, вес внутренних органов, вес внешних тканей и количество колец.

Из исходной выборки случайным образом было взято 500 векторов, образовавших контрольную выборку (КВ), оставшиеся 3677 векторов образовали эталонную выборку (ЭВ). Для формирования трех видов тестовых выборок (ТВ) в ЭВ случайным образом были созданы пропуски в 10% элементов: ТВ с пропусками в категорных данных (КТВ), ТВ с пропусками в непрерывных данных (НТВ) и ТВ смешанного типа (СТВ), образованные в результате наложения пропусков из НТВ и КТВ на ЭВ. Далее в соответствии с методом АСК была осуществлена процедура восстановления пропусков в трех ТВ с получением ВКТВ, ВНТВ и ВСТВ соответственно. Вопрос качественного анализа замен пропущенным данным, формируемых кластеризационным алгоритмом, освещен в работе [17] и выходит за рамки настоящего исследования.

Для анализа влияния выбросов на результат работы нейросетевых классификаторов была создана ШЭВ посредством добавления в ЭВ 150 векторов, равномерно распределенных по всему диапазону изменения значений атрибутов в ЭВ случайным образом, представляющих «шум». Под шумом понимаются случайные данные, представляющие нерелевантную информацию. По аналогии с ЭВ на базе ШЭВ были созданы ШВКТВ, ШВНТВ и ШВСТВ.

С целью преодолеть эффект переобучения для многослойных персептронов каждая ТВ разделялась на два подмножества: основное (90% векторов) и контрольное (10% векторов).

Ко времени работы РБФ и МП добавлялись затраты на восстановление информационных потерь в соответствии с АСК (30-32 сек.) для всех выборок, кроме ЭВ и ШЭВ.

Результаты и их обсуждение. Полученные временные (в сек.) показатели работы нейроклассификаторов приведены в табл. 1.

Таблица 1

Тип ИНС	Тестовая выборка				Тестовая выборка с «шумом»			
	ВКТВ	ВНТВ	ВСТВ	ЭВ	ШВКТВ	ШВНТВ	ШВСТВ	ШЭВ
МП	205	241	164	122	191	233	235	162
РБФ	77	76	76	45	79	80	79	49
СХН	73	88	71	43	79	77	84	52
КСХН	42	43	42	42	43	44	43	44

Экспериментальные результаты свидетельствуют о значительном, в 1,7-5,4 раза, превосходстве КСХН над альтернативными архитектурами по скорости обработки данных с информационными пропусками за счет тесной интеграции механизмов восстановления данных с алгоритмом кластеризации. В то же время следует отметить сопоставимую с СХН и РБФ скорость обучения, с разницей в 2-18%, при работе с эталонными данными. Относительная нестабильность показателей СХН обусловливается особенностями работы алгоритма самоорганизации, большое влияние на который оказывает начальный выбор центров кластеров [14].

Несмотря на хорошие скоростные результаты, показанные СХН на 5 выборках из 8, КСХН обучается в среднем в 1,65 раза быстрее. Близость временных показателей у СХН и КСХН на эталонных выборках обусловливается необходимостью расчета КСХН элементов матрицы смежности и БОС вне зависимости от наличия пропусков данных.

Показатели СКО работы ИНС на КВ, приведенные в табл. 2, свидетельствуют о незначительном (13%) совокупном росте качества работы КСХН по сравнению с СХН.

Таблица 2

Тип ИНС	Тестовые выборки				Тестовые выборки с «шумом»			
	ВКТВ	ВНТВ	ВСТВ	ЭВ	ШВКТВ	ШВНТВ	ШВСТВ	ШЭВ
МП	0,021	0,031	0,025	0,020	0,030	0,035	0,034	0,047
РБФ	0,024	0,023	0,027	0,022	0,034	0,041	0,048	0,056
СХН	0,027	0,028	0,031	0,024	0,036	0,042	0,048	0,059
КСХН	0,023	0,025	0,028	0,021	0,036	0,039	0,038	0,051

Подобный эффект объясняется менее равномерным распределением объектов по кластерам для КСХН, нежели для СХН и РБФ, в основе которых лежит процесс самоорганизации. Наблюдаемое среднее 10% превосходство МП над КСХН объясняется, очевидно, более сложной архитектурой МП.

Отдельно стоит коснуться вопроса работы с выборками, содержащими шум. Здесь на трех выборках из четырех КСХН продемонстрировала 14% и 17% рост точности по сравнению с РБФ и СХН соответственно. Такой эффект может быть объяснен лучшей работой алгоритма кластеризации, осуществляющего первичное формирование и последующую установку центров кластеров на основе плотностного подхода. При повышении плотности шума качество обучения КСХН снизится до уровня СХН ввиду того, что посторонний сигнал начнет образовывать плотные регионы, рассматриваемые АСК как самостоятельные кластеры.

Влияние выбора параметров селективного алгоритма на качество (в смысле СКО $Ч10^2$) работы КСХН на примере обработки ВСТВ отражено в табл. 3. Компромиссным, с точки зрения скорости работы и качества, является установка параметра A=4 при выборке 7% элементов из кластера (B = 7).

Таблица 3

Объем селективных подмножеств B (%)	Количество селективных подмножеств (A)							
	2	3	4	5	6	7	8	9
5	3,241	2,919	2,897	2,885	2,885	2,851	2,842	2,836
7	2,917	2,864	**2,844**	2,843	2,830	2,831	2,822	2,821
10	2,920	2,855	2,840	2,827	2,822	2,824	2,822	2,818
13	2,888	2,860	2,840	2,828	2,819	2,819	2,820	2,816

Сравнительный анализ временных и качественных показателей работы бисекционного выборочного алгоритма с альтернативными подходами на ВСТВ и ШВСТВ приведен в табл. 4. Здесь различные методы вторичной кластеризации были применены к кластерам, сформированным алгоритмом АСК.

Таблица 4

Алгоритм	СКО $Ч10^2$		Время (сек.)	
	ВСТВ	ШВСТВ	ВСТВ	ШВСТВ
Кохонена	3,072	4,396	21,6	27,8
K-вероятностный	3,327	4,159	18,8	19,0
J-средних	2,776	3,762	16,6	17,9
Бисекционный	2,811	3,875	17,5	19,1
Бисекционный выборочный	2,844	3,837	12,3	13,5

Классический самоорганизационный алгоритм (Кохонена) продемонстрировал наименее удовлетворительные результаты, опередив лишь *K*-вероятностный подход на ВСТВ по величине СКО. Разница в 29% между временем работы алгоритма самоорганизации на двух выборках почти равного размера наглядно подтверждает нестабильность процесса сходимости. *K*-вероятностный [29] алгоритм также не может быть рекомендован из-за высокой ошибки работы.

Наилучшей точности установки центров кластеров удалось достичь посредством алгоритма *J*-средних [30], превзошедшего результаты бисекционного выборочного подхода всего на 2% по величине СКО, при этом на треть уступившего ему в скорости работы.

Сравнение бисекционного и бисекционного выборочного методов позволяет выявить особенности выборочного механизма: применение в расчетах части информации несколько снижает точность установки центров кластеров на выборке ВСТВ, обусловливая в то же время меньшую чувствительность к шуму на выборке ШВСТВ. Поскольку разница по величине СКО у обоих методов не превосходит 1%, более важным показателем является скорость работы, где выборочный метод демонстрирует 40% преимущество.

Заключение

Описанная в настоящей работе система позволила решить комплекс проблем, характерных для искусственных нейронных сетей Хехт-Нильсона, что подтверждается экспериментальными результатами, полученными на открытой БД.

В частности, было достигнуто совокупное 1,7- 2-кратное ускорение работы СХН при незначительном повышении точности за счет перехода от алгоритмов самоорганизации к двухуровневому методу кластеризации. Преодолена проблема информационной неполноты обучающей выборки посредством применения методов прямых и групповых замен пропущенных элементов из наиболее близких векторов в кластерах. Частично решена проблема выбросов, влиявших на качество аппроксимации в конечных кластерах.

К недостаткам предложенного подхода следует отнести прежде всего высокие накладные расходы на расчет и хранение данных о межвекторных расстояниях и информации о ближайших общих соседях. Другим ограничением следует считать проблему подбора оптимального количества нейронов Кохонена в скрытом слое СХН, фактически представляющую собой задачу оптимального разбиения сложного пространственного кластера. Наконец, открытой остается проблема неглубоких локальных минимумов, характерная для метода бисекции из раздела 2.2, в связи с чем дальнейшее совершенствование системы связано, по нашему мнению, с разработкой новой версии бисекционного метода кластеризации на основе алгоритма EM [27].

ЛИТЕРАТУРА

1. *Головко В.А*. Нейронные сети: обучение, организация и применение. Кн.4. – М.: ИПРЖР, 2001.
2. *Hecht-Nielson R*. Counterpropagation Networks // Proceedings of the IEEE First International

Conference on Neural Networks. – N.Y., 1987. – P.19-32.
3. *Тархов Д.А*. Нейронные сети. Модели и алгоритмы. Кн. 18. – М.: Радиотехника, 2005.
4. *Логинов Б.М., Аюев В.В*. Нейросетевые агенты в задачах управления с разделенными по времени входными данными высокой размерности // Нейрокомпьютеры: разработка, применение. – 2007. – № 5. – С. 31-41.
5. *Kaufman L., Rousseeuw P*. Finding Groups in Data – An Introduction to Cluster Analysis. – NY: John Wiley & Sons, 1990.
6. *Ssali G., Tshilidzi M*. Computational intelligence and decision trees for missing data estimation // Proceedings of the IEEE International Joint Conference on Neural Networks. – Hong Kong, 2008. – P. 201-207.
7. *Nelwamondo F.V., Mohamed S., Marwala T*. Missing Data: A comparison of neural network and expectation maximization techniques // Current Science. – 2007. – Vol. 3, N11. – P. 1514-1521.
8. *Wasito I., Mirkin B*. Nearest neighbors in least-squares data imputation algorithms with different missing patterns // Computational Statistics & Data Analysis. – 2006. – Vol. 5, N4. – P. 926-949.
9. *Local* linear regression for generalized linear models with missing data / Wang C.Y., Wang S., Carroll R.J., Gutierrez R.G. // Annals of Statistics. – 1998. – N3(26). – P. 1028-1050.
10. *Fujikawa Y., Ho T*. Cluster-based algorithms for dealing with missing values // Advances in Knowledge Discovery and Data Mining: Lecture Notes in Computer Science. – NY: Springer, 2002. – Vol. 2336. – P. 549-554.
11. *Wu X., Barbara D*. Learning missing values from summary constraints // ACM SIGKDD Exploration Newsletters. – 2002. – Vol. 4, N1. – P. 21-30.
12. *Метод* доменной компенсации информационной неполноты БД / Аюев В.В., Аунг З.Е., Тура А., Логинова М.Б. // Труды МГТУ им. Н.Э.Баумана. – М., 2007. – Т. 594. – С. 57-64.
13. *Gan G., Ma C., Wu J*. Data Clustering: Theory, Algorithms, and Applications. – Philadelphia: SIAM Press, 2007.
14. *Kohonen T*. Self-Organizing Maps. – Berlin: Springer, 1995.
15. *Haykin S*. Neural Networks: A Comprehensive Foundation. 2-nd Edition. – New Jersey: Prentice-Hall, 1999.
16. *Аюев В.В., Тура А., Логинова М.Б.* Модификация метода статической кластеризации для работы с несвязными данными // Труды МГТУ им. Н.Э.Баумана. – М., 2007. – Т. 594. – С. 64-71.
17. *Метод быстрой динамической кластеризации неоднородных данных* / Аюев В.В., Тура А., Лайнг Н.Н., Логинова М.Б. // Системы управления и информационные технологии. – 2008. – № 3(33). – С. 26-29.
18. *Карпухин П.А., Логинова М.Б., Аюев В.В.* Алгоритм двухуровневой кластеризации // Тр. Всерос. конф. «Наукоемкие технологии в приборо- и машиностроении и развитие инновационной деятельности в вузе» / МГТУ им. Н.Э.Баумана. – М., 2008. – Т. 2. – С.99-104.
19. *A* density-based algorithm for discovering clusters in large spatial databases with noise / Ester M., Kriegel H.P., Jцrg S., Xu X. // Proc. of 2nd International Conference on Knowledge Discovery and Data Mining. – LA, 1996. – P. 226-231.
20. *Steinbach P.N., Kumar M., Tan V*. Introduction to Data Mining. International Edition. – NY.: Addison Wesley, 2006.
21. *Ertoz L., Steinback M., Kumar V*. Finding clusters of different sizes, shapes, and density in noisy, high dimensional data // Proceedings of Second SIAM International Conference on Data Mining. – San Francisco, 2003. – P. 47-58.
22. *Hinneburg A., Keim D.A.* A general approach to clustering in large databases with noise // Knowledge and Information Systems. – 2003. – N. 5. – P. 387-415.
23. *Wang L., Wang Z.O.* CUBN: A clustering algorithm based on density and distance // International Conference on Machine Learning and Cybernetics. – 2003. – Vol. 1. – P. 108-112.
24. *Guha S., Rastogi R., Shim K.* CURE: An efficient clustering algorithm for large databases // Proceedings of ACM SIGMOD, 1998. – Santa Clara, 1999. – P. 73-84.

25. *When* is Nearest Neighbors Meaningful? / Beyer K., Goldstein J., Ramakrishnan R., Shaft U. // Proc. of the Int. Conf. Database Theory. – N.Y., 1999. – P. 217-235.
26. *Mitchell T.M.* Machine Learning. – NY: McGraw-Hill, 1997.
27. *Hogg R., McKean J., Craig A*. Introduction to Mathematical Statistics. – NJ: Pearson Prentice Hall, 2005.
28. *Asuncion A*., *Newman D.J.* UCI Machine Learning Repository. – Irvine: University of California, School of Information and Computer Science, 2008.
29. *Wishard D.* K-means clustering with outlier detection, mixed variables and missing values // In Schwainger M, Opitz O. Explorary data analysis in empirical research. – N.Y.: Springer, 2002. – P. 216-226.
30. *Hansen P., Mladenovic N.* J-means: A local search heuristic for minimum sum of squares clustering // Pattern Recognition, 2001. – N34(2). – P.405-413.

Статья представлена к публикации членом редколлегии Ю.А. Григорьевым.

E-mail:
Аюев В.В. – Vadim.Ayuyev@gmail.com.

УДК 004.855

© 2009 г. **А.С. Клещев**, д-р физ.-мат. наук,
С.В. Смагин
(Институт автоматики и процессов управления ДВО РАН, Владивосток)

НЕКОТОРЫЕ СВОЙСТВА МЕТОДА СЛУЧАЙНОЙ РАССТАНОВКИ ГРАНИЦ ПЕРИОДОВ ДИНАМИКИ[1]

Приведена постановка задачи индуктивного формирования знаний для упрощенной онтологии медицинской диагностики, а также представлены результаты экспериментального исследования метода случайной расстановки границ периодов динамики для этой онтологии.

Ключевые слова: индуктивное формирование знаний, метод индуктивного формирования знаний, модельные данные, онтология медицинской диагностики, экспериментальное исследование свойств метода.

Введение

В работе [1] представлен обзор существующих экспериментальных исследований в области изучения свойств методов индуктивного формирования знаний (ИФЗ), а также приведена постановка задачи таких исследований. В работах [1, 2]

[1] Работа выполнена при финансовой поддержке ДВО РАН в рамках Программы №2 фундаментальных исследований Президиума РАН “Интеллектуальные информационные технологии, математическое моделирование, системный анализ и автоматизация”, проект “Развитие систем управления базами знаний с коллективным доступом”.