УДК 004.9

В. И. Аникин, д-р техн. наук, проф., e-mail: anikin_vi@mail.ru,
Поволжский государственный университет сервиса, г. Тольятти,
А. А. Карманова, инженер-программист, e-mail: turaeva.alexandra@mail.com,
ООО "НетКрэкер", г. Тольятти

Кластеризация и классификация многомерных данных клеточной нейронной сетью Кохонена

Показана перспективность и высокая временная эффективность кластеризации и классификации многомерных выборок данных нейронной сетью Кохонена, обучаемой клеточным автоматом. Продемонстрировано полезное применение краевого эффекта и многосвязных самоорганизующихся карт Кохонена для решения проблемы "мертвых" нейронов и надежного выделения границ группировки кластеров в пространстве с линейно и/или нелинейно разделимыми классами учебных образцов.

Ключевые слова: *нейронная сеть Кохонена, клеточный автомат, Excel, классификация многомерных данных, визуализация, U-матрица, P-матрица*

Введение

Ввиду огромного количества информации, которая, согласно оценкам аналитиков, удваивается каждые 2—3 года, лишь очень малая ее часть будет когда-либо увидена человеческим глазом. Единственная возможность понять и найти что-то полезное в ней — использовать методы *Data Mining (Knowledge Discovery in Data — KDD*). Технология KDD позволяет решать задачи многомерной кластеризации, регрессии, поиска ассоциаций, классификации и используется во многих областях с большим объемом данных — астрономии, биологии, медицине, телекоммуникациях, банковском деле, промышленном производстве и т. д. Важная и быстрорастущая часть KDD — анализ связей между данными, имеющий приложения в биоинформатике, поисковых системах и др.

Существующие парадигмы исследовательского анализа многомерных данных базируются на трех аспектах: 1) предобработка в целях приведения данных к форме, удобной для последующего анализа; 2) кластеризация данных, т. е. их группировка в физически осмысленные наборы/множества; 3) визуализация полученных результатов таким образом, чтобы сходство и различие данных было хорошо видно.

Достоинством самоорганизующихся карт Кохонена (*Self-Organizing Map — SOM*) является то, что они позволяют решать задачу анализа данных в едином процессе, объединяющем все три перечисленных аспекта [1]. Действительно, самоорганизующаяся карта Кохонена выполняет нелинейное отображение многомерных данных на одно-, двух- или трехмерную решетку и поэтому представляет собой эффективный инструмент визуализации связей между данными. Метод SOM применим к любым данным, которые можно представить векторами их свойств.

В ранних работах по методу SOM использовали для классификации небольшие одномерные и двумерные нейронные сети (НС) Кохонена в режиме "один нейрон на один кластер/класс" (*k-means* SOM) [1]. Если число узлов SOM отличается от ожидаемого числа классов, то требуются группировка или разделение кластеров в целях выделения необходимого числа классов.

Важным методом группирования кластеров в методе SOM является визуализация пространственного распределения нейронов с помощью *U*-матрицы Ультша (визуализация, основанная на расстояниях) вида, показанного в табл. 1 [2], где используются четыре типа расстояний от центра нейрона u_{ij} до его ближайших соседей: $dx(i, j) = d(u_{ij}, u_{i+1, j})$, $dy(i, j) = d(u_{ij}, u_{i, j+1})$, $dxy(i, j) = d(u_{ij}, u_{i+1, j+1})$, $dyx(i, j) = d(u_{i+1, j}, u_{i, j+1})$ и $dz(i, j) = 0{,}5(dxy(i, j) + dyx(i, j))$. Элементы $du(i, j)$ матрицы могут быть произвольными. В нашем исследовании значения этих последних элементов равны числу учебных образцов, попавших в окрестность Вороного нейрона u_{ij}, в совокупности эти элементы образуют так называемую *H*-матрицу. Достоинством табличной реализации *UH*-матрицы является то, что инструмент условного форматирования Excel дает возможность визуализировать границы между классами учебных образцов интерактивно, путем варьирования значения порогового расстояния, отделяющего соседние классы друг от друга.

Другим методом группирования кластеров, также предложенным Ультшем [3], является метод *P*-матрицы. *P*-матрица имеет размерность *U*-матрицы, а значения ее элементов равны числу учебных образцов внутри гиперсфер заданного радиуса, проведенных вокруг нейронов и промежуточных аппроксимирующих точек. В дополнение к информации о расстояниях, предоставляемой *U*-матрицей, *P*-матрица содержит информацию о плотности распределения учебных образцов во входном пространстве (визуализация, основанная на плотности).

Таблица 1

U-матрица	...	$2j-1$	$2j$	$2j+1$	...
...	...	...	...	...	...
$2i-1$	...	$dz(i-1, j-1)$	$dy(i-1, j)$	$dz(i-1, j)$	...
$2i$	...	$dx(i, j-1)$	$du(i, j)$	$dx(i, j)$	...
$2i+1$	...	$dz(i, j-1)$	$dy(i, j)$	$dz(i, j)$	...
...	...	...	...	...	...

Таблица 2

P-матрица	...	$2j-1$	$2j$	$2j+1$	...
...	...	...	...	...	...
$2i-1$	...	$w_{\text{ср}}(i-1, j-1)$	$w_{\text{ср}}(i-1, j)$	$w_{\text{ср}}(i-1, j)$	...
$2i$	...	$w_{\text{ср}}(i, j-1)$	$w(i, j)$	$w_{\text{ср}}(i, j)$	...
$2i+1$	...	$w_{\text{ср}}(i, j-1)$	$w_{\text{ср}}(i, j)$	$w_{\text{ср}}(i, j)$	...
...	...	...	...	...	...

В нашем исследовании координаты аппроксимирующих точек для *P*-матрицы определялись как показано в табл. 2, где $w_{\text{ср}}(i, j \pm 1)$ — среднее значение одноименных координат соседних нейронов по горизонтали SOM; $w_{\text{ср}}(i \pm 1, j)$ — по вертикали; $w_{\text{ср}}(i \pm 1, j \pm 1)$ — по диагоналям.

Цель нашего исследования — повышение эффективности алгоритма обучения и качества классификации, достигаемые за счет использования клеточной НС Кохонена с многосвязной SOM. Клеточная нейронная сеть Кохонена всегда обучается на максимальной скорости в пакетном режиме, что позволяет сократить время обучения в несколько десятков раз по сравнению с классическим алгоритмом обучения [1]. О насущной необходимости повышения временной эффективности алгоритма обучения НС Кохонена классическим алгоритмом говорят, например, условия экспериментов по классификации генов, выполненных в эффектной работе [4]: размерность данных — 79, число генов в выборке — 2460, число итераций обучения — $3{,}5 \cdot 10^6$, время обучения на рабочей станции с процессором Alpha — 19 ч, число реализаций опыта — практически неограниченно.

Обоснование и постановка задачи исследования

Важной особенностью данного исследования в сравнении с работой [5] является то, что здесь оптимизирован алгоритм адаптивного саморазворачивания нейронной сети, выполняемый в первых эпохах обучения. Концептуально главная идея описанного в работе [5] алгоритма саморазворачивания сети — постепенное втягивание активными нейронами неактивных нейронов в нормированный гиперкуб пространства учебных образцов, — осталась неизменной, но использованный в этой работе алгоритм саморазворачивания претерпел три изменения.

1. В первой эпохе обучения в качестве нейрона-победителя выбирается центральный, а не угловой нейрон SOM, благодаря чему число эпох i_p саморазворачивания сети по сравнению с алгоритмом в работе [5] сократилось вдвое и равно $i_p = \lceil (L-1)/2 \rceil$, где L — длина большей стороны прямоугольной SOM, скобки $\lceil\ \rceil$ означают округление вверх до ближайшего целого.

2. Мультипликативная поправка (dw_x, dw_y) к весам соседей нейрона-победителя вычисляется по формуле $(dw_x, dw_y) = 1 - \alpha(i, j)$, учитывающей структуру обучающего клеточного автомата (КА) Мура, где (i, j) — смещение вновь активируемых нейронов относительно их активных соседей в плоскости саморазворачивания сети xy; $\alpha = 10^{-3}...10^{-4}$ — масштабный коэффициент. Относительные смещения (i, j) активируемых нейронов определяются из простых соображений: в первой эпохе обучения (рис. 1) центральный нейрон-победитель с координатами (0, 0) втянет 8 нейронов из окружения радиуса $r = 1$ в точки плоскости xy с относительными смещениями (−1, −1), (0, −1), (1, −1), (−1, 0), (1, 0), (−1, 1), (0, 1), (1, 1), во второй эпохе обучения эти 8 нейронов, в свою очередь, втянут в нормированный гиперкуб пространства образцов своих соседей из окружения радиуса $r = 2$ и поместят их в точки со смещениями (−2, −2), (−2, −1), ..., (2, 1), (2, 2) и т. д.

3. К двум статическим типам связей между нейронами (⊞ — связь включена, ⊟ — связь выключена), определенным в работе [5], добавлен еще один, адаптивный тип связи [6]: ⊞ — связь включена в эпохах саморазворачивания сети и автоматически отключается в последующих эпохах обучения. Благодаря этому типу связей саморазворачивание сети всегда выполняется клеточным автоматом Мура, причем во всем входном пространстве учебных образцов, рассматриваемом как единое целое, кроме того упрощается процедура начальной инициализации клеточных НС Кохонена с многосвязными SOM [5].

Выбор плоскости xy, в которой происходит саморазворачивание сети, неоднозначен (всего имеется $d(d-1)/2$ таких плоскостей, где d — размерность пространства учебных образцов) и оказывает влияние на результат кластеризации данных методом SOM. Очевидно, что оптимальной, но не необходимой, является плоскость двух главных компонент [7].

Рис. 1. Адаптивная активация нейронов за три эпохи саморазворачивания клеточной НС Кохонена

В описываемых ниже экспериментах по кластеризации и классификации данных клеточной НС Кохонена использовались тестовые моделируемые компьютером и взятые из архива *UCI Machine Learning Repository* экспериментальные выборки многомерных данных:

1. Моделируемая выборка данных объемом 2000 элементов с равномерным случайным распределением учебных образцов в двух соприкасающихся углами кубах; трехмерный вариант выборки, предложенной Ултшем [3] для обоснования необходимости введения *P*-матрицы.

2. Моделируемая трехмерная случайная выборка данных Chainlink объемом 1000 элементов, предложена Ултшем [2] для обоснования необходимости введения *U*-матрицы.

3. 13-мерная экспериментальная выборка данных *Wine recognition* объемом 178 элементов для классификации производителей вина: http://mlr.cs.umass.edu/ml/datasets/Wine.

4. Трехмерная экспериментальная выборка данных Haberman's Survival объемом 306 элементов для классификации послеоперационной выживаемости больных раком: URL: https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival.

Кластеризацию элементов тестовых выборок данных и их классификацию методом группирования проводили посредством созданной в Excel табличной модели клеточной НС Кохонена. Визуализацию результатов кластеризации выполняли с использованием встроенных средств форматирования Excel, она включала автоматическое построение по результатам кластеризации интерактивных *UH*- и *P*-матриц, координатных *C*-карт, карты классов учебных образцов (при обучении с учителем).

Метод решения задачи исследования

Для решения поставленной задачи исследования использовали две клеточные НС Кохонена, реализованные в виде итерационных табличных моделей Excel [8].

1. Число нейронов в сети — 49, размер прямоугольной SOM — 7 × 7.

2. Число нейронов в сети — 400, размер прямоугольной SOM — 20 × 20.

Число нейронов в каждой сети и размеры SOM можно произвольно изменять в сторону уменьшения путем деактивации некоторых нейронов.

Заметим, что использование табличных моделей НС Кохонена двух размеров не является принципиально необходимым, но позволяет уменьшить время, затрачиваемое на кластеризацию многомерных выборок данных в Excel. Поэтому на обучающих выборках данных небольшого размера, таких как выборки Фишера [5] и Хабермана, использовали табличную модель НС Кохонена размером 49 нейронов, а на выборках большого размера (объемом в тысячи образцов) — размером 400 нейронов (размер сети выбирается из соображений, чтобы каждый кластер в среднем содержал единицы—десятки учебных образцов).

Обе табличные модели клеточной НС Кохонена были реализованы одинаково на пяти рабочих листах Excel: "Выборки", "Инициализация", "Обучение", "Кластеризация", "Визуализация".

Рабочий лист "Выборки" содержит приведенные выше многомерные выборки данных, нормированные к интервалу [−1, 1] по каждой числовой координате, а также интерфейс активации, используемой в данный момент выборки.

Рабочий лист "Инициализация" предназначен для задания начальной конфигурации клеточной НС Кохонена, позволяя пользователю определить геометрию нейронной сети и поперечные размеры прямоугольной SOM, явно указать нейрон, который станет победителем в первой эпохе обучения. На этом же рабочем листе выполняются расчеты мультипликативных поправок (dw_x, dw_y) к весам нейронов в эпохах саморазворачивания сети и вычисляются координаты случайной точки на гиперсфере начальной инициализации весов нейронов [5].

На рабочем листе "Обучение" табличными средствами реализован описанный в работе [5] оригинальный алгоритм обучения клеточной НС Кохонена. Пользовательский интерфейс этого листа позволяет в визуальном режиме определять параметры обучения (число эпох саморазворачивания, взаимодействия, дообучения по алгоритму WTA), разрывать/активировать связи между нейронами, запускать/останавливать итерации обучения.

На рабочем листе "Кластеризация" реализован алгоритм выполнения одной эпохи подачи учебных образцов на входы обученной НС в целях их подразделения на кластеры и вычисления элементов *P*-матрицы для заданного радиуса гиперсфер, окружающих нейроны и промежуточные точки.

Наконец, рабочий лист "Визуализация" содержит интерактивные средства визуализации результатов кластеризации. Используя встроенные в Excel формулы и средства условного форматирования, удается эффективно визуализировать *UH*- и *P*-матрицы, координатные *C*-карты, карты классов образцов (при обучении с учителем), а также интерактивно выполнять разнообразную пост-обработку полученных результатов.

С нашей точки зрения, именно погруженность клеточной НС Кохонена в мощную вычислительную и аналитическую среду Excel является главным достоинством данного экспериментального исследования, открывая перед исследователем потрясающие возможности для креативного творчества и визуального экспериментирования с алгоритмом вычислений, входными, выходными и промежуточными данными.

Работа с табличной моделью клеточной НС Кохонена включает следующие шаги (рис. 2).

1. В зависимости от обучающей выборки данных выбирается один из двух файлов Excel: для вы-

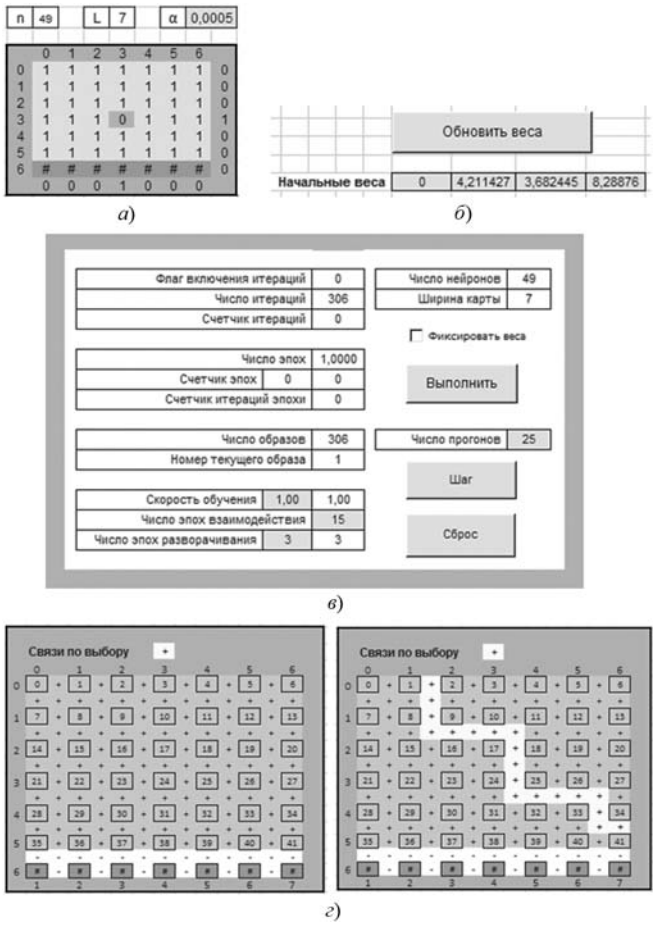

Рис. 2. Пользовательский интерфейс табличной модели НС Кохонена с прямоугольной SOM размером 6 × 7 нейронов:
а — задание структуры сети; *б* — выбор случайной точки на гиперсфере инициализации весов нейронов; *в* — пользовательский интерфейс алгоритма обучения НС; *г* — типы связей между нейронами SOM: слева — конфигурация Мура, справа — двухсвязная конфигурация

борок объемом до 500—1000 образцов — табличная модель клеточной НС Кохонена размером 7 × 7 нейронов, для выборок большего объема — размером 20 × 20 нейронов.

2. С помощью пользовательского интерфейса, показанного на рис. 2, *а*, где активные нейроны помечены символом 1, нейрон-победитель — символом 0, незадействованные нейроны — символом #, определяется структура клеточной НС Кохонена, указывается нейрон-победитель в первой эпохе обучения, задается значение масштабного коэффициента α.

3. Щелчком на кнопке Обновить веса выбирается случайная начальная точка на гиперсфере инициализации начальных весов нейронов (рис. 2, *б*).

4. На рабочем листе *Обучение* задаются параметры обучения: скорость обучения; число эпох саморазворачивания сети; взаимодействия, обучения/прогонов (рис. 2, *в*).

5. Указываются типы связей между нейронами НС Кохонена в эпохах взаимодействия (рис. 2, *г*).

6. Запускается процесс обучения сети (кнопка Выполнить на рис. 2, *в*).

7. На рабочем листе *Кластеризация* выполняется эпоха кластеризации элементов выборки данных обученной НС Кохонена.

8. На рабочем листе *Визуализация* выполняется интерактивный анализ результатов кластеризации, в случае необходимости изменяются структура и параметры нейронной сети, и шаги 3—8 повторяются заново.

Результаты экспериментального исследования

Действенность описанного алгоритма адаптивного саморазворачивания клеточной НС Кохонена продемонстрируем на примере двумерной случайной выборки данных с круговой симметрией объемом 4000 учебных образцов (рис. 3).

Видно, что вновь активируемые нейроны втягиваются в нормированное пространство учебных образцов латерально, без скручивания сети. Краевой эффект, обычно считающийся паразитным, теперь играет положительную роль, оставляя на периферии пространства учебных образцов свободное место для добавления следующих активируемых нейронов.

Эксперименты по кластеризации и классификации элементов тестовых выборок данных состояли в следующем. Сначала НС Кохонена с односвязной SOM обучалась клеточным автоматом Мура (рис. 4, *а*), выполнялась кластеризация элементов обучающей выборки и делалась попытка группировки полученных кластеров. Для этого, варьируя порог группирования, посредством *UH*- и *P*-матриц наблюдали изменения вида границы между клас-

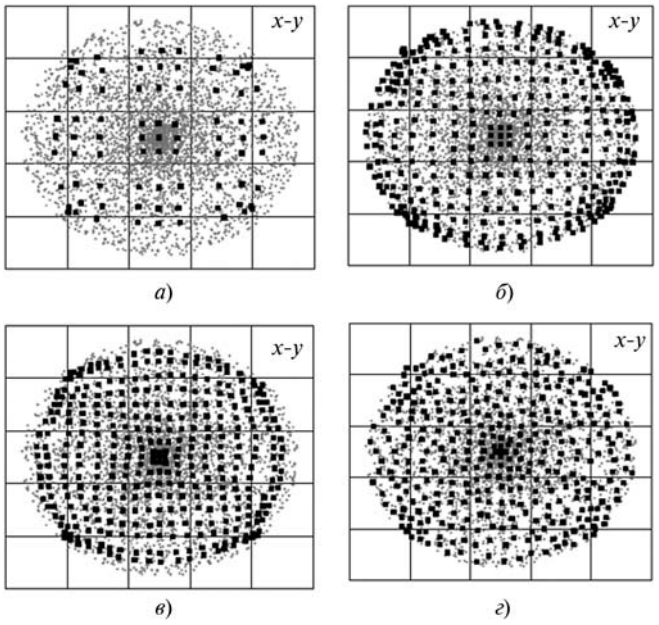

Рис. 3. Обучение нейронной сети Кохонена с прямоугольной SOM размером 19 × 19 нейронов:
а — 4 эпохи саморазворачивания; *б* — 9 эпох саморазворачивания сети; *в* — 15 эпох взаимодействия; *г* — 20 эпох обучения

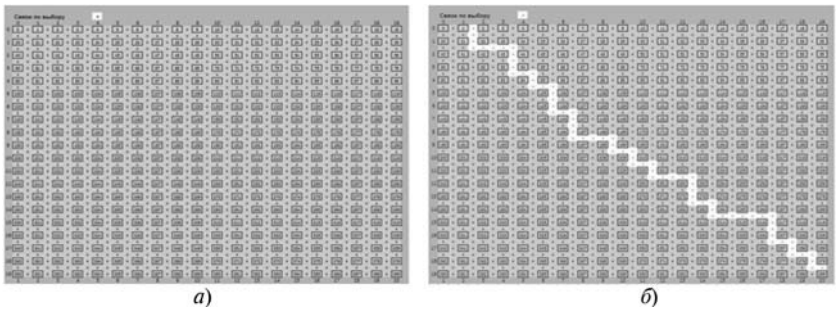

Рис. 4. Обучающие клеточные автоматы для случайной выборки данных 1

сами (обычно она выглядит "размытой", нечетко выраженной).

Затем на базе полученных результатов группирования, односвязная SOM разрезалась на несколько областей вдоль предполагаемых границ классов (рис. 4, *б*), и НС Кохонена с полученной многосвязной прямоугольной SOM обучалась вторично. Для достижения более качественной группировки, операцию разрезания карты SOM иногда приходилось выполнять несколько раз. После завершения процесса обучения снова выполнялись кластеризация элементов обучающей выборки и группировка кластеров.

Физические закономерности кластеризации и классификации многомерных данных исследовались на приведенных выше тестовых выборках данных 1 и 2.

Примечательной особенностью тестовой обучающей выборки данных 1 является то, что два класса ее образцов пространственно линейно разделимы, но у соприкасающихся вершин кубов образцы, принадлежащие разным классам, отстоят друг от друга на небольшом расстоянии, и при кластеризации могут быть ошибочно отнесены к другому классу.

Параметры обучения НС Кохонена выборкой данных 1: число нейронов в сети — 400; размер SOM — 20 × 20, начальные веса нейронов — (0; 7,754; −6,273; 0,727), нейрон-победитель в первой эпохе обучения — 189; начальные веса нейрона-победителя — (0; 0; 0; 0), число эпох саморазворачивания сети — 10; число эпох взаимодействия — 30; всего эпох обучения — 40.

Результаты классификации элементов выборки 1 при заданных параметрах обучения позволяют сделать ряд важных выводов (рис. 5, 6).

При классификации элементов выборки данных клеточной НС Кохонена с односвязной SOM наблюдаются около 20 "мертвых" нейронов, не победивших ни разу (рис. 5, *а* слева). Появление "мертвых" нейронов обусловлено тем, что смежные с ними нейроны относятся к разным классам, поэтому в эпохах взаимодействия эти "мертвые" нейроны, сами находясь вне области концентрации учебных образцов, благодаря связям со своими со-

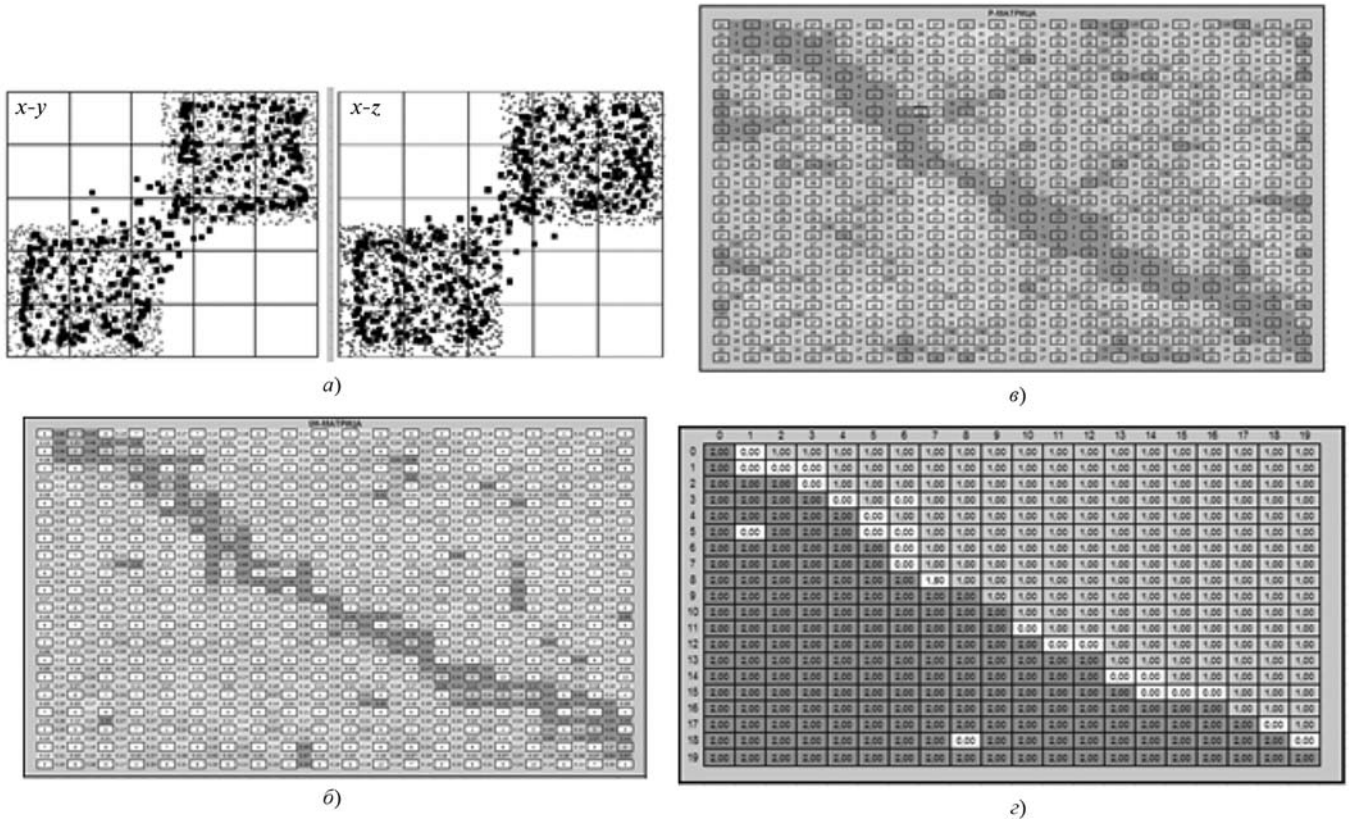

Рис. 5. Результаты кластеризации элементов выборки данных 1 после 30 эпох обучения, односвязная SOM:
а — 30 эпох обучения; *б* — *UH*-матрица: порог = 0,32; *в* — *P*-матрица: *R* = 0,2, порог = 20; *г* — карта классов

седями поочередно подтягиваются к победившим образцам то одного, то другого класса. При использовании двухсвязной SOM связи между нейронами, принадлежащими разным классам, принудительно разрываются, и появление "мертвых" нейронов исключается (рис. 6, *а*).

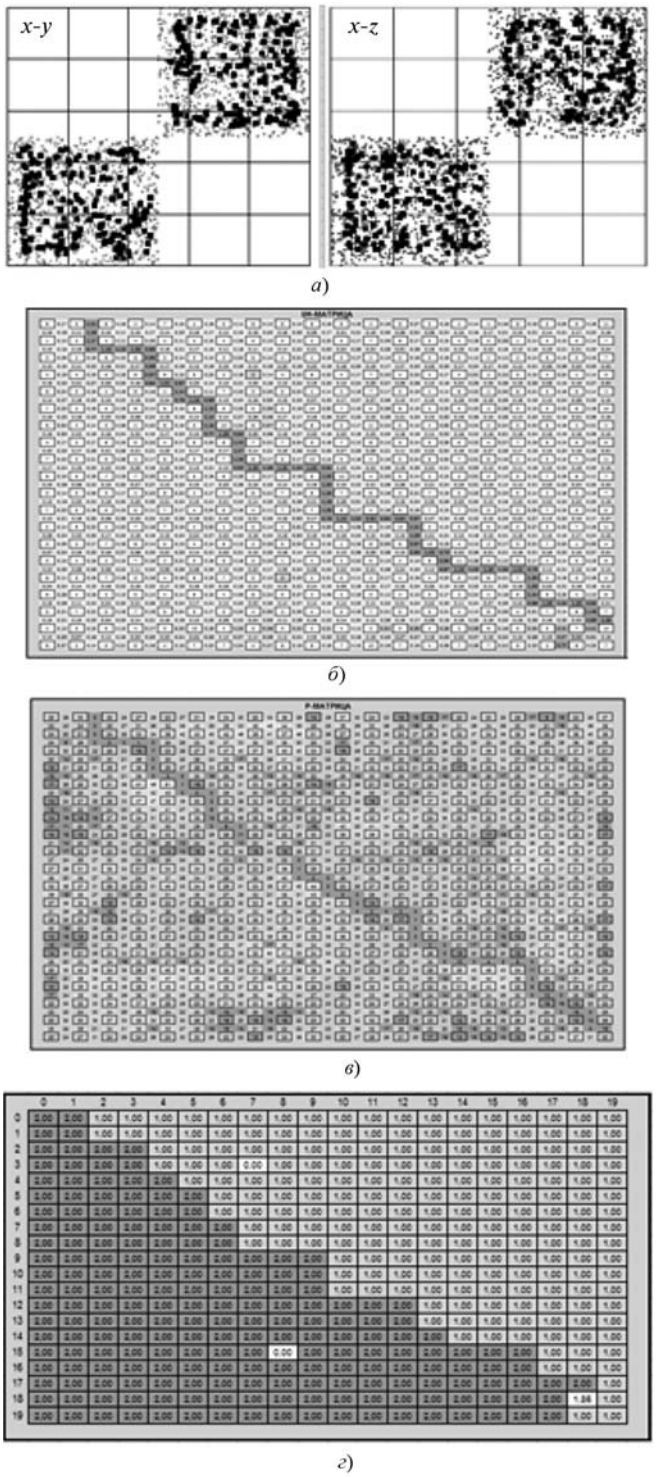

Рис. 6. Результаты кластеризации элементов выборки данных 1 после 30 эпох обучения, двухсвязная SOM:
а — 30 эпох обучения; *б* — *UH*-матрица: порог = 0,62; *в* — *P*-матрица: *R* = 0,2, порог = 20; *г* — карта классов

Краевой эффект, с которым обычно борются путем задания периодических граничных условий на краях SOM, наблюдается и у границ пространственно разделимых классов образцов. На самом деле этот эффект внутренне присущ самому методу SOM, и избавиться от него невозможно, пока существуют латеральные локальные связи между нейронами. Единственный корректный механизм борьбы с ним — это дообучение (после завершения эпох взаимодействия) нейронной сети Кохонена по алгоритму WTA.

В случае многосвязной SOM улучшается разделение учебных образцов на классы посредством *UH*-матрицы, граница между классами становится четко определенной (рис. 6, *б*), повышается порог группирования, что равносильно увеличению расстояния между граничными элементами разных классов. Положительный эффект от использования многосвязной SOM в случае *P*-матрицы менее заметен (рис. 5, *в*, 6, *в*).

Аналогичные закономерности кластеризации и классификации данных клеточной НС Кохонена выявляют эксперименты с тестовой обучающей выборкой 2. Элементы этой выборки распределены в трехмерном пространстве так, что они образуют два тонких сцепленных друг с другом подобно олимпийским кольцам тора, один из которых лежит в плоскости *xy*, второй — в плоскости *xz* (рис. 7, *а*).

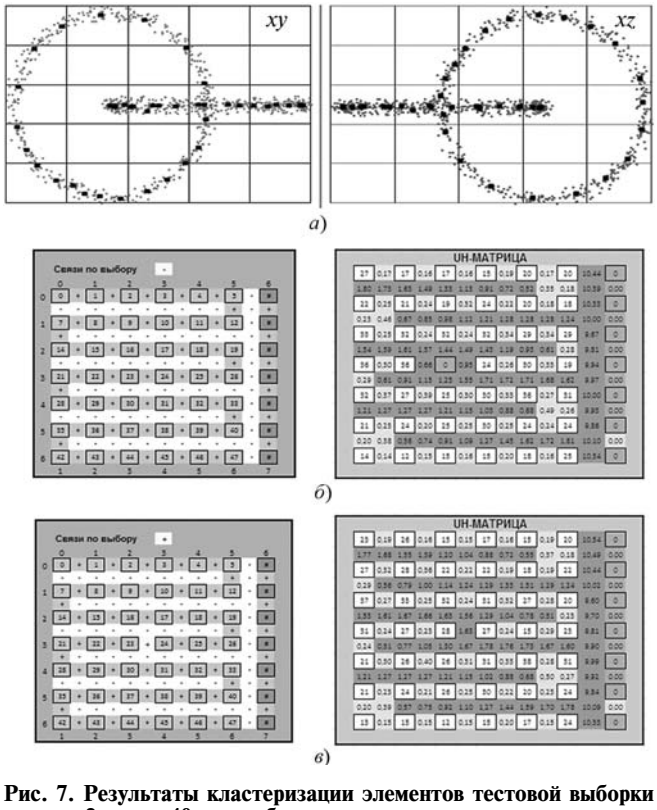

Рис. 7. Результаты кластеризации элементов тестовой выборки данных 2 после 40 эпох обучения:
а — проекции координат образцов на плоскости *xy* и *xz*; *б* — односвязная линейная SOM и *UH*-матрица при пороге 0,50; *в* — двухсвязная линейная SOM и *UH*-матрица при пороге 0,50

Отличительной особенностью данной выборки является то, что ее элементы линейно не разделимы в пространстве, хотя и образуют два явно выраженных класса.

Попытка классифицировать элементы выборки данных двухклеточной НС Кохонена с двумерной SOM к успеху не привела, и нам пришлось воспользоваться НС Кохонена с линейной SOM (рис. 7, *б*). Это означает, что структура нейронной сети, применяемой для классификации многомерных данных, зависит от их пространственного распределения, поэтому в общем случае успешное практическое применение метода SOM предполагает наличие предварительных знаний об этом распределении.

Следует также отметить, что число эпох саморазворачивания НС Кохонена с линейной SOM довольно большое и равно половине длины последней. Это позволяет наглядно наблюдать, как постепенно, по мере выполнения эпох саморазворачивания, сеть начинает все лучше понимать пространственную структуру данных. По завершении процесса обучения нейроны распределяются в пространстве образцов так, как показано на рис. 7, *а*.

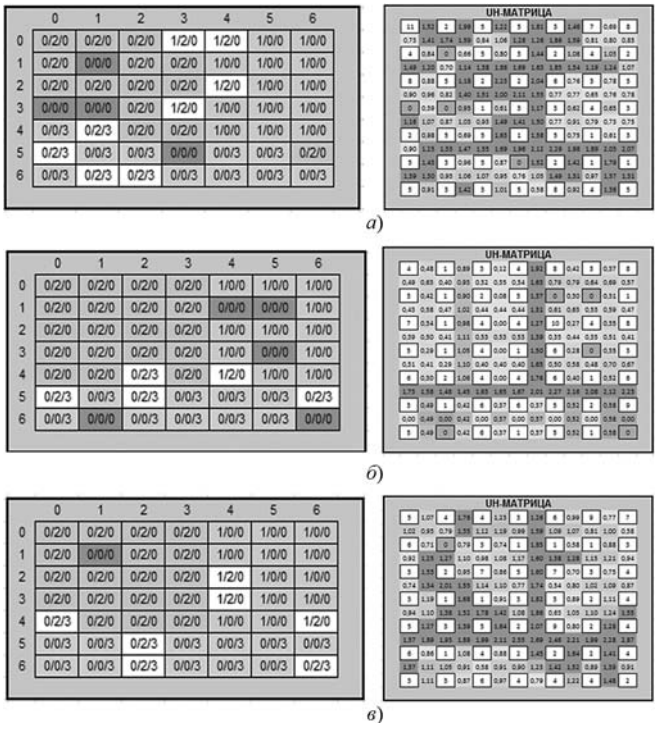

Рис. 8. Результаты кластеризации элементов экспериментальной выборки данных *Wine recognition*:
а — односвязная SOM: 40 эпох обучения, порог = 1,25; *б* — трехсвязная SOM, 30 эпох взаимодействия, порог = 1,25; *в* — трехсвязная SOM: 40 эпох обучения, порог = 1,25

C	0	1	2	3	4	5	6
0	1,42	1,11	1,50	1,23	1,09	1,20	1,50
1	1,33	1,11	1,00	1,00	1,00	1,33	2,00
2	1,33	1,17	1,11	1,25	1,00	1,67	1,57
3	1,43	1,07	1,57	1,20	1,21	1,67	1,50
4	1,33	1,67	1,50	1,11	1,50	1,33	1,50
5	1,09	1,00	1,11	1,38	1,25	1,20	1,50
6	0	0	0	0	0	0	0

Рис. 9. Карта классов выборки данных Хабермана: порог выделения — 1,25

Как и в экспериментах с выборкой данных 1, разрыв связи между нейронами, расположенными у границы двух классов, устраняет причину появления "мертвых" нейронов, граница между классами на *UH*-матрице становится четко выделенной, заметно повышается порог группирования (рис. 7, *в*).

Такие же закономерности наблюдаются и при классификации образцов экспериментальной выборки данных *Wine recognition*, только в этом случае прямоугольную SOM приходится разрезать на три области по числу ожидаемых классов (рис. 8).

Форма контура разрезания определяется из вида *UH*- и *P*-матриц, а также карты классов. Метки на карте классов показывают, образцы каких классов содержатся в каждом кластере, например, метка 0/2/3 означает, что в соответствующем кластере имеются учебные образцы второго и третьего классов.

Заметим также, что, как показывает более внимательный анализ *UH*-матрицы, образцы класса 2 делятся по средней вертикальной линии (рис. 8, *б*) на два вложенных субкласса. О существовании этих субклассов в описании экспериментальной выборки данных *Wine Recognition* ничего не говорится.

Экспериментальная выборка данных Хабермана является особой и отличается от предыдущих тестовых выборок тем, что классифицировать ее элементы методом группирования кластеров не удается. Действительно, варьирование значений порогов группирования для *UH*- и *P*-матриц этой выборки не позволяет выявить четкую границу между классами больных, выживших и умерших от рака. Причиной этому является то, что элементы выборки данных Хабермана не являются пространственно разделимыми (ни линейно, ни нелинейно).

Решить задачу классификации элементов этой выборки данных клеточной НС Кохонена можно, тем не менее, методом обучения с учителем. На рис. 9 показана карта классов, полученная после обучения этой выборкой клеточной НС Кохонена с прямоугольной SOM размером 6 × 7 нейронов при следующих параметрах: начальные веса нейронов — (0; 1,547; 6,117; 7,512); координаты нейрона-победителя 24 — (0; 0; 0; 0); эпох разворачивания сети — 3; эпох взаимодействия — 15; эпох обучения — 25.

Светлой заливкой на карте выделены нейроны с долей больных, которые умерли от рака в течение пяти лет после операции, не превышающей 25 %.

Заключение

Проведенные экспериментальные исследования показали перспективность и высокую временную эффективность авторского алгоритма кластеризации

и классификации многомерных выборок данных клеточной НС Кохонена с многосвязной SOM. Исследования проводили с использованием итерационной табличной модели этой сети, реализованной в Excel чисто табличными средствами, без написания программного кода VBA. Временная сложность алгоритма обучения сети равняется $T = O(n \cdot N \cdot d \cdot i)$, где n — число нейронов в сети; N — объем обучающей выборки; d — размерность входного пространства; i — число эпох обучения. Экспериментальное время выполнения одной эпохи обучения для $n = 400$, $N = 8000$, $d = 4$ составило $t \approx 100$ с, т. е. алгоритм обучения клеточной НС Кохонена работает в Excel приблизительно в 2 раза медленнее, чем классический алгоритм обучения НС Кохонена в программе SOM_PAK [1]. Это незначительное возрастание времени обучения клеточной НС Кохонена, очевидно обусловленное низким быстродействием электронных таблиц, многократно окупается наглядностью, гибкостью анализа и эффективной визуализацией результатов кластеризации и классификации многомерных данных в Excel.

Показано полезное применение краевого эффекта и многосвязных SOM для надежного выделения границ группирования кластеров в линейно и нелинейно разделимых пространствах учебных образцов и для решения проблемы "мертвых" нейронов, расположенных вблизи этих границ.

Список литературы

1. **Кохонен Т.** Самоорганизующиеся карты: пер. 3-го англ. изд. М.: БИНОМ. Лаборатория знаний, 2008. С. 158—337.
2. **Ultsch A., Siemon H. P.** Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis // Proc. INNC'90 Neural Networks Conf. 1990. P. 305—308.
3. **Ultsch A.** U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data // Department of Computer Science, University of Marburg, Technical Report. 2003. N. 36. P. 1—12.
4. **Nikkilä J., Törönen P., Kaski S., Venna J., Castrén E., Wong G.** Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps // Neural Networks. 2002. N. 15. P. 953—966.
5. **Аникин В. И., Карманова А. А.** Обучение искусственной нейронной сети Кохонена клеточным автоматом // Информационные технологии. 2014. № 11. С. 73—80.
6. **Shah-Hosseini H., Safabakhsh H.** A TASOM-based Algorithm for Active Contour Modeling // Pattern Recognition Letters. 2003. V. 24. P. 1361—1373.
7. **Marghescu D.** Multidimensional Data Visualization Techniques for Financial Performance Data: A Review // TUCS Technical Report N. 810, 2007. P. 1—32.
8. **Аникин В. И., Аникина О. В.** Визуальное табличное моделирование клеточных автоматов в Microsoft Excel. Тольятти: Изд-во ПВГУС, 2013. 321 с.

V. I. Anikin, Professor, e-mail: anikin_vi@mail.ru, Volga State University of Service, Tolyatti,
A. A. Karmanova, Software Engineer, e-mail: turaeva.alexandra@mail.com, LLC "NetCracker", Tolyatti

Clustering and Classification of Multidimensional Data by Kohonen's Cellular Neural Network

The paper presents experimental results of multidimensional data clustering and classification by help of Kohonen's cellular neural network (CNN). An important feature of our study is that Kohonen's CNN and visualization tools (self-organizing map (SOM), U-, H- and P-matrixes, coordinate maps, map of data classes) have been implemented in Microsoft Excel spreadsheet, without programming in VBA. The user interface of this spreadsheet model makes it easy to change the configurable parameters and visually observe the neural network learning process and data classification results using cluster grouping method.

The experimental studies have shown high temporal efficiency of Kohonen's CNN learning algorithm. Having the time complexity of the algorithm is $T = O(n \cdot N \cdot d \cdot i)$, where n — the number of neurons, N — the size of training sample, d — the input space dimension, i — the number of training epochs, the experimental run time of one training epoch was $t \approx 100$ s for $n = 400$, $N = 8000$, $d = 4$. That is, the learning algorithm of Kohonen's CNN works in Excel about 2 times slower than the classic learning algorithm in the SOM_PAK program. This slight training time increase in Excel is repeatedly paid by clarity, analysis flexibility and effective visualization of multi-dimensional data clustering patterns.

A useful application of the edge effect and multilinked SOMs for reliable identification of clusters grouping boundaries in linear and nonlinear separable input spaces and the possibility of solving a well-known problem of "dead" neurons are shown.

Keywords: *Kohonen's neural network, cellular automata, Excel, multi-dimensional data classification, visualization, U-matrix, P-matrix*

References

1. **Kohonen T.** Samoorganizujuschiecija karti. Moskva: BINOM. Laboratoriya znanii. 2008. P. 159—337.
2. **Ultsch A., Siemon H. P.** Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. *Proc. INNC'90 Neural Networks Conf.* 1990. P. 305—308.
3. **Ultsch A.** U*-Matrix: A Tool to Visualize Clusters in High Dimensional Data. Department of Computer Science, University of Marburg, Technical Report. 2003. N. 36. P. 1—12.
4. **Nikkilä J., Törönen P., Kaski S., Venna J., Castrén E., Wong G.** Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps. *Neural Networks.* 2002. N. 15. P. 953—966.
5. **Anikin V. I., Karmanova A. A.** Obuchenie iskusstvennoi neironnoi seti Kohonena kletochnim avtomatom. *Informatcionnie technologii*. 2014. N. 11. P. 73—80.
6. **Shah-Hosseini H., Safabakhsh H.** A TASOM-based Algorithm for Active Contour Modeling. *Pattern Recognition Letters.* 2003. V. 24. P. 1361—1373.
7. **Marghescu D.** Multidimensional Data Visualization Techniques for Financial Performance Data: A Review. *TUCS Technical Report N. 810*, 2007. P. 1—32.
8. **Anikin V. I., Anikina O. V.** *Vizualnoe tablichnoe modelirovanie kletochnih avtomatov v Mjocrosoft Excel.* Tolyatti: PVGUS, 2013. 321 p.