УДК 004.032.26(043)

ЭКСПЕРИМЕТАЛЬНЫЕ ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ КАРТ КОХОНЕНА ДЛЯ ВИЗУАЛИЗАЦИИ ЗАКОНОМЕРНОСТЕЙ

Н.С. Лесная[1], В.Б. Репка[2], О.В. Ивченко[3], А.В. Шерстнюк[4]

[1]ХНУРЭ, г. Харьков, Украина, lmd@kture.kharkov.ua
[2] ХНУРЭ, г. Харьков, Украина, victoria_repka@kture.kharkov.ua
[3]ХНУРЭ, г. Харьков, Украина, ivchenko.o@mail.ru
[4]ХНУРЭ, г. Харьков, Украина, troll_andrey@mail.ru

Проведён анализ способов визуализации карт Кохонена для выявления скрытых структур в данных низкой и высокой размерности. Оценена эффективность нескольких способов визуализации карт Кохонена. Доработаны и программно реализованы способы визуализации карт Кохонена. Исследована возможность применения нейронной сети Кохонена и построенных по результатам её работы карт для сегментации двумерных и многомерных входных векторов. На основании анализа экспериментов сделан вывод о целесообразности использования определённых способов визуализации при решении задачи сегментации.

НЕЙРОННАЯ СЕТЬ, КАРТА КОХОНЕНА, ВИЗУАЛИЗАЦИЯ, МАТРИЦА РАССТОЯНИЙ, ДИАГРАММА ХИНТОНА, КЛАСТЕРИЗАЦИЯ

Введение

Визуализация данных – задача, с которой сталкивается в своей работе любой исследователь. К задаче визуализации данных сводится проблема представления в наглядной форме данных эксперимента или результатов теоретического исследования. Традиционные инструменты в этой области – графики и диаграммы – плохо справляются с задачей визуализации, когда возникает необходимость изобразить более трёх взаимосвязанных величин [1].

На сегодняшний день одним из самых используемых способов визуализации являются самоорганизующиеся карты Кохонена, с помощью которых возможно эффективно визуально отобразить скрытые закономерности в исследуемых данных. Существует множество способов визуализации карт Кохонена, таких как унифицированная матрица расстояний, проекция Саммона, матрица плотности попадания, матрица кластеров, матрица ошибок квантования и другие, каждый из которых позволяет наблюдать те или иные свойства анализируемых данных.

Из-за большого разнообразия способов визуализации становится актуальной задача экспериментального исследования эффективности их работы на разных наборах данных, с разной размерностью и с разными структурами. Таким образом, ставится задача исследования различных методов визуализации карты Кохонена при анализе картой различных выборок, для определения степени применимости конкретных способов визуализации в конкретной рассматриваемой задаче.

1. Особенности нейронных сетей Кохонена (двумерный случай)

Нейронная сеть Кохонена (самоорганизующаяся карта/Self-organizing map/SOM) относится к самоорганизующимся сетям, которые при поступлении входных сигналов, в отличие от сетей, использующих обучение с учителем, не получают информации о желаемом выходном сигнале. В связи с этим невозможно сформировать критерий настройки, основанный на рассогласовании реальных и требуемых выходных сигналов ИНС. Все предъявляемые входные сигналы из заданного обучающего множества самоорганизующаяся сеть в процессе обучения разделяет на классы, строя так называемые топологические карты [3].

Существует несколько алгоритмов обучения сети Кохонена, отличающиеся способом определения выигравших нейронов, а также различные модификации этих алгоритмов. В работе использован алгоритм обучения, описанный в [4]. После окончания процесса обучения карта Кохонена классифицирует входные примеры по группам схожести, что можно отобразить цветами на карте.

Для визуализации карт Кохонена могут быть использованы 1-, 2- и 3-мерные пространства, но обычно практически ограничиваются отображением с помощью 2-мерных поверхностей, так как именно в таком виде человек воспринимает геометрические структуры наиболее естественно, и отношения между объектами выглядят наиболее наглядно.

Под визуализацией данных картой Кохонена понимается такой способ представления многомерного распределения данных на двумерной плоскости, при котором качественно отражены основные закономерности, присущие исходному распределению: его кластерная структура, внутренние зависимости между признаками, информация о расположении данных в исходном пространстве и другие. Карта признаков, полученная алгоритмом SOM, является топологически упорядоченной в том смысле, что пространственное положение нейронов в решётке соответствует конкретной

области или признаку входного образа [4]. Две точки, близко лежащие на карте Кохонена, будут близки и в *N*-мерном входном пространстве, но не наоборот.

В результате обучения сети Кохонена строится совокупность карт, каждая из которых представляет двумерную сетку узлов, размещенных в многомерном пространстве. Самый простой вариант — использование градаций серого цвета. Также для раскраски можно использовать любую иную градиентную палитру.

Рассмотрим наиболее наглядные варианты визуализации карт Кохонена на примере с 2-мя входными характеристиками и многомерный случай. Для проведения экспериментов была использована выборка «Spirals» (рис. 1), которая представляет собой две спирали (два линейно неразделимых класса) и состоит из 2-мерного вектора входных данных и 194-х базовых точек. Набор данных «Spirals» был выбран для описания визуализации карт Кохонена, так как его отображение на картах даёт наглядные и простые для понимания результаты.

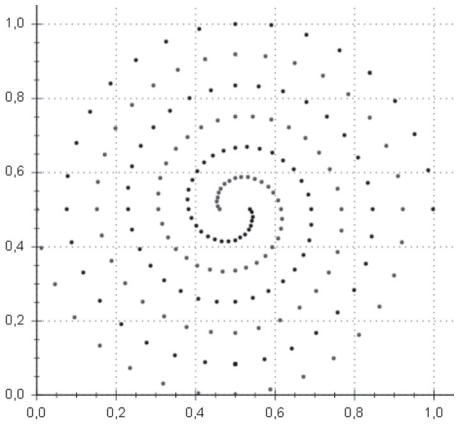


Рис. 1. Выборка «Spirals»

На данной выборке рассмотрим применение карт Кохонена для решения задачи кластеризации, то есть представим, что число классов заранее неизвестно. Приведем результаты, полученные за 100 проходов нейронной сети Кохонена. В качестве функции соседства выбрана функция Гаусса, радиус в начале обучения – 3, норма обучения – 0,5.

2. Способ визуализации посредством унифицированной матрицы расстояний

Унифицированная матрица расстояний (U-matrix) – это представление SOM, визуализирующее расстояние между нейронами. Рассчитывается расстояние между соседними нейронами и представляется в виде различной раскраски между соседними вершинами. Темная раскраска соответствует большему расстоянию между нейронами и соответственно – промежуткам между величинами кодовой книги (codebook) в пространстве входных векторов. Светлая раскраска между нейронами означает, что векторы кодовой книги (codebook vectors) расположены близко друг к другу в пространстве входных векторов [5]. Светлые области могут считаться кластерами, а темные – разделителями между ними. Это может быть полезным представлением при попытке найти кластеры во входных данных без наличия априорной информации о кластерах.

Для построения матрицы расстояний необходимо определить расстояние между весовыми коэффициентами нейрона и его ближайшими соседями, как показано на рис. 2.

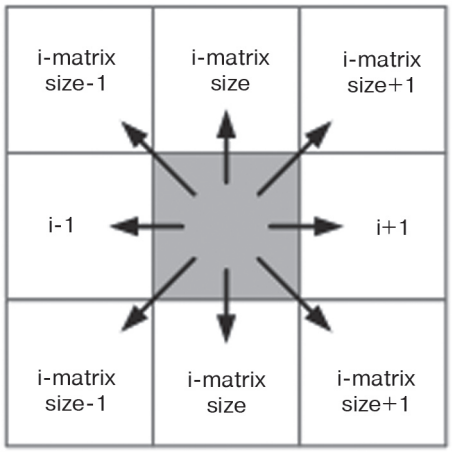


Рис.2. Ближайшие соседи *i*-го нейрона

В качестве меры близости был выбран квадрат евклидова расстояния (1):

$$d_{pq} = \sum_{i}^{n} (w_{pi} - w_{qi})^2 , \qquad (1)$$

где d_{pq} – квадрат евклидова расстояния между нейронами p и q; w_{pi} – i-й весовой коэффициент нейрона p; w_{qi} – i-й весовой коэффициент нейрона q; n – число весовых коэффициентов нейрона.

Оттенок ячейки, расположенной между двумя узлами, отражает расстояние между узлами в исходном пространстве. Оттенок самого узла вычисляется с помощью усреднения. Оттенки раскраски карты задаются в соответствии с формулами 2-4:

$$R' = 255 \cdot \left(\frac{R - R_{\min}}{R_{\max} - R_{\min}} \right), \qquad (2)$$

$$G' = 255 \cdot \left(\frac{G - G_{\min}}{G_{\max} - G_{\min}} \right), \qquad (3)$$

$$B' = 255 \cdot \left(\frac{B - B_{\min}}{B_{\max} - B_{\min}} \right), \qquad (4)$$

где R, G, B – составляющие цветовой модели, а min и max – минимальное и максимальное расстояния на карте Кохонена.

На рис. 3 приведена матрица расстояний, отображенная в оттенках серого, которая соответствует карте размером 15×15 (225 нейронов). SOM полностью покрывает все точки выборки, и 194 узла

на карте соответствуют расположениям данных точек. Светлые участки соответствуют точкам, находящимся близко друг от друга; в центре изображения чётко видны две спирали, по мере удаления от центра, спирали становятся нечёткими, так как увеличивается расстояние между точками, и ближайшими соседями для самых дальних точек будут ближайшие точки другой спирали. Можно сделать вывод, что входная выборка представляет собой два класса, границы которых чётко видны в центре изображения.

Карта, матрица расстояний для которой приведена на рис. 3, построена за 500 проходов алгоритма обучения, приемлемый результат достигается уже на 100-м проходе – видны спирали, но изображение не ровное.

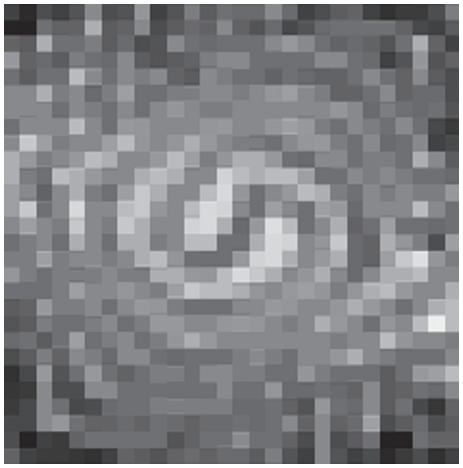
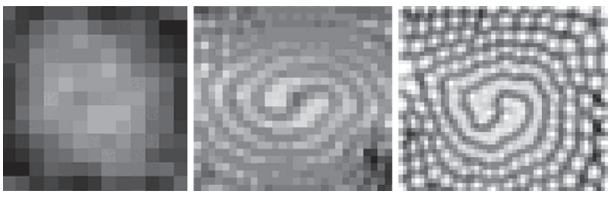


Рис. 3. Матрица расстояний

На рис. 4 приведены матрицы расстояний для карт разного размера: 7×7, 15×15 и 25×25. Похоже выглядят U-матрицы для карт размером 15×15, 30×30 и 50×50, если из их алгоритма раскраски убрать ячейки, расположенные между узлами, которые отображают расстояния между ними. Если карта больше выборки, ячейки карты будут заполнены неактивными нейронами, которые будут заменять эти ячейки с подобными оттенками, если меньше – то оттенки узлов будут отображать расстояния. Визуализация, реализованная подобным образом, является более чёткой и может быть использована в случаях, когда входная выборка приемлемого размера и увеличение размера карты в 2 раза незначительно скажется на скорости обучения нейронной сети.

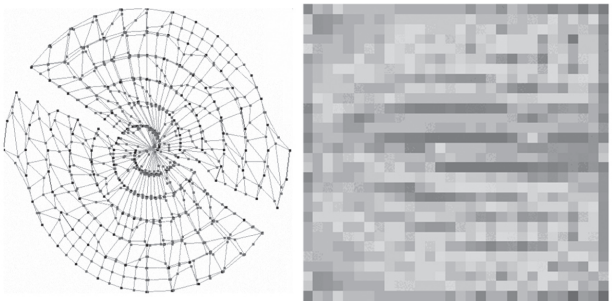


а *б* *в*

Рис. 4. Матрицы расстояний для карт Кохонена разного размера: *а* – 7 × 7; *б* – 15 × 15; *в* – 25 × 25

Точность визуализации и дальнейшего анализа зависят от размера карты. Это обусловлено тем, что во время переноса точек из пространства на карту каждой точке сопоставляется ближайший узел карты. Это приводит к тому, что средний квадрат расстояний от точки до её проекции на карте сильно зависит от количества узлов в сетке. На первой карте (рис. 4*а*), сетка которой состоит из наименьшего числа узлов, видно, что в центре изображения точки расположены близко друг другу; чем дальше от центра — расположение становится более разреженным, что также характерно и для входного вектора. Размер последней карты (рис. 4*в*) слишком велик, и в отдельные кластеры начинают отделяться дальние точки спиралей, что обусловлено в данной ситуации избытком неактивных нейронов.

Одной из проблем построения карт Кохонена является топологический дефект. Если изначально параметр расстояния σ выбран малым или очень быстро уменьшается, то далеко расположенные нейроны не могут влиять друг на друга [2]. Результатом этого является топологический дефект карты Кохонена, приведённый на рис. 5*а*.



а *б*

Рис. 5. Топологический дефект карты Кохонена:
а – дефект карты Кохонена;
б – влияние дефекта на *U*-Matrix

Как видно на рис. 5 *а*, обе части карты настроились правильно, но на ней явно виден дефект, который отрицательно сказывается на матрице расстояний, построенной по данной карте (рис. 5 *б*). Такая карта, в зависимости от степени дефекта, не может быть интерпретирована корректно.

3. Визуализация диаграммой Хинтона

Популярным способом визуализации является диаграмма Хинтона. На каждом узле сетки изображается квадрат, размер которого пропорционален числу точек, ближайших к данному узлу, а оттенок соответствует значению соответствующего отображаемого признака [1].

На рис. 6 приведена диаграмма Хинтона для карты размером 7×7. Данная сетка представляет собой карту, подходящую для отображения двумерного и многомерного пространства данных.

Чем больше точек попадает в узел, тем больше соответствующий размер квадрата. Оттенок квадрата соответствует значению цвета на матрице расстояний для данного нейрона.

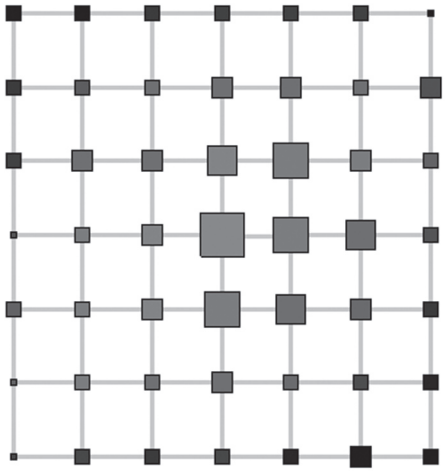


Рис. 6. Диаграмма Хинтона для «Spirals»

На рис. 7 приведена сетка карты Кохонена, соответствующая диаграмме Хинтона, и исходный набор данных, который она покрыла, то есть точки исходного набора были распределены по узлам карты.

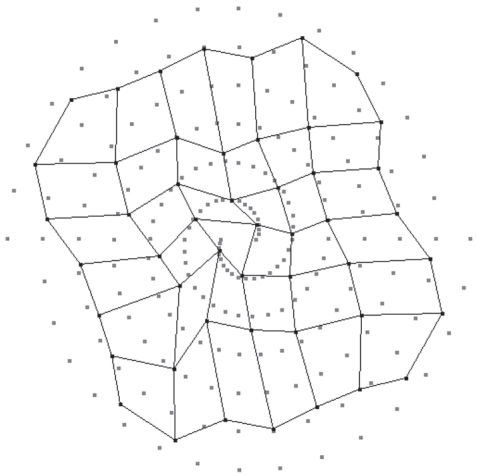


Рис. 7. Карта Кохонена 7×7

В выборке «Spirals», приведённой на рис. 1, в центре наблюдается сгущение точек, поэтому в центре диаграммы Хинтона находятся узлы большего размера, а по краям – меньшего.

Также частоту попадания точек в узел можно отобразить на каждой эпохе, как показано на рис. 8.

Карты, показанные на рис. 8, соответствуют картам частот. Они построены в соответствии с количеством реагирований нейронов-победителей на данные из обучающего множества на разных эпохах. На данных картах размер круга соответствует числу реагирований нейрона за эпоху.

Диаграмма Хинтона может соответствовать матрице плотности попадания.

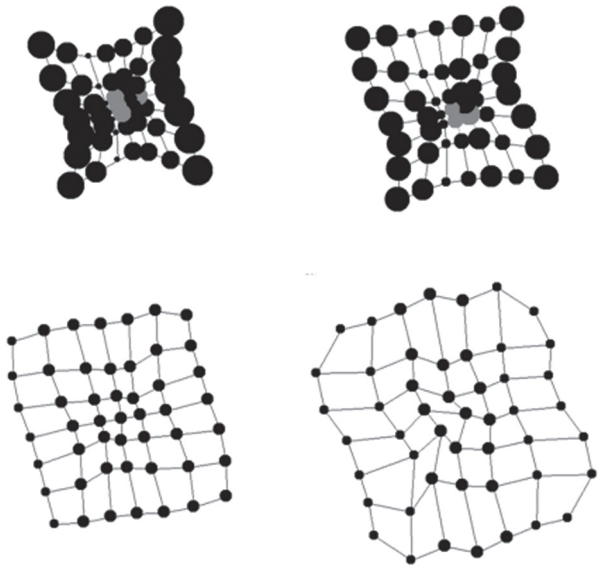


Рис. 8. Карты Кохонена 7×7 на разных эпохах

4. Применение карт Кохонена для многомерной кластеризации

Рассмотрим применение карт Кохонена для выявления скрытых закономерностей в данных высокой размерности на примере решения задачи кластеризации кредитных историй и информации о заемщиках.

Набор исходных данных состоит из 150 образов. Каждый образ представляет собой информацию о заемщиках и выданных кредитах и состоит из 31 характеристики (сумма кредита, срок кредита, цель кредитования, возраст, пол, образование, квартира и другое).

Рассмотрим визуализацию карт Кохонена и определение кластеров на выборке из 146 образов, после чего классифицируем 4 образа с помощью обученной НС.

На рис. 9 и рис. 10 показаны матрицы расстояний и соответствующие им диаграммы Хинтона.

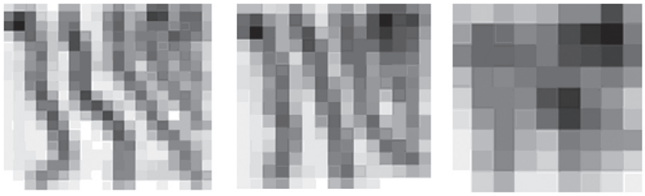


Рис. 9. Матрицы расстояний для карт разного размера

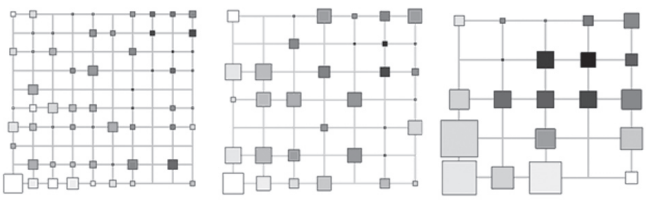


Рис. 10. Диаграммы Хинтона для карт разного размера

По картам на рис. 9 и 10 можно сделать следующие выводы. На картах большего размера видны 3-4 области сгущения точек, отображённые светлым цветом, каждая из которых плавно переходит в другую, они соответствуют областям в *N*-мерном входном пространстве.

На всех диаграммах, приведённых на рис. 10, есть нейроны, которые ни разу не среагировали на данные из обучающей выборки, в то же время, другие нейроны среагировали несколько раз, что говорит о том, что в исходном обучающем множестве есть очень схожие образы. В табл. 1 приведены результаты реагирования нейронов на обучающую выборку для карт разного размера.

Таблица 1

Размер карты	% среагировавших нейронов	% не среагировавших нейронов
10×10	60	40
7×7	69,4	30,6
5×5	84	16

По результатам, приведенным в табл. 1, видно, что число не среагировавших нейронов с уменьшением размеров карты уменьшается незначительно, пока карта не достигает слишком малого размера, например, 25 нейронов.

Узлы диаграммы Хинтона также можно представить в виде кластеров, тогда образы, попавшие в один узел, будут соответствовать одному классу. Ни рис. 11 приведена диаграмма Хинтона, представляющая собой 4 узла, то есть отображающая 4 класса. Оттенки узлов соответствуют степени удаленности их друг от друга, а также сгущению точек, то есть самый удалённый узел карты также соответствует и наименьшему числу образов, попавшим в него.

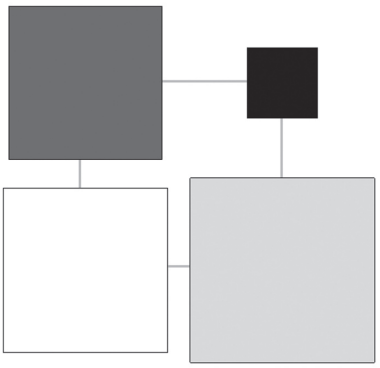


Рис. 11. Диаграмма Хинтона для карты Кохонена с 4-мя нейронами

В табл. 2 приведено число попаданий в каждый узел для диаграммы, приведённой на рис. 11.

В табл. 3 приведены наиболее весомые усреднённые значения по результатам сегментации нейронной сетью Кохонена информации о заемщиках в банках. По рис. 11 и табл. 2 и 3 можно сделать следующие выводы: самым небольшим по числу принятых образов является четвёртый сегмент, он удалён от трёх других кластеров и точки, находящиеся в нём, разрежены. Кроме того, в нём находится наименьшее количество образов. По результатам табл. 3 видна аналогичная ситуация: это лица в возрасте свыше 60 лет – пенсионеры, что вполне соответствует небольшому числу таких заемщиков, у которых есть множество расхождений в характеристиках, и эти характеристики почти не пересекаются с характеристиками других классов.

Таблица 2

Номер узла	Цвет	Число отнесённых образов
1	Тёмно-серый	39
2	Светло-серый	47
3	Белый	42
4	Чёрный	18

Таблица 3

Характеристики	Усреднённые значения свойств сегментов			
	Сегмент 1	Сегмент 2	Сегмент 3	Сегмент 4
Возраст	до 22 лет	от 22 до 40	от 40 до 60	свыше 60
Семейное положение	холост (не замужем)	женат (замужем)	женат (замужем)	разведён(а), вдовец(а), женат (замужем)
Количество иждивенцев	0	2-3	3-5	до 2
Социальный статус	студент	не руководящий работник, руководящий работник	руководящий работник	пенсионер
Образование	среднее, неоконченное высшее	высшее, специальное	высшее	среднее, высшее
Занятость	нет	да	да	нет
Квартира	нет	да	да	да
Машина	нет	нет	да	да, нет
Среднемесячный доход	до 500	от 1500	от 2000	до 500
Среднемесячный расход	до 400	от 1000	от 1000	до 400
Основное направление расходов	образование, аренда недвижимости, оплата услуг	покупка товаров длительного пользования	покупка и ремонт недвиж, покупка товаров длител. пользования	оплата услуг
Цель кредита	образование, потребительские товары, иное	потребительские товары, транспорт, недвижимость	транспорт, недвижимость	потребительские товары, иное

Наиболее приближен к нему первый – тёмно серый сегмент – лица в возрасте до 22 лет. Приближение данных сегментов обусловлено значи-

тельным отдалением от других сегментов, также у них есть схожие характеристики, отличные от двух других сегментов, такие как: небольшой среднемесячный доход и расход, незанятость. Данные сегменты не очень схожи друг с другом, их приближение обусловлено значительным отдалением от других сегментов. К первому сегменту менее всего приближен третий сегмент, характеристики этих сегментов значительно отличаются.

Самые крупные по числу отнесённых к ним образов второй – светло серый и третий – белый сегменты, расположенные наиболее близко друг к другу, это лица в возрасте от 22 до 40 лет и от 40 до 60 соответственно. Также по результатам, приведенным в табл. 3, эти сегменты содержат образы с наиболее схожими характеристиками, такими как семейное положение, количество иждивенцев, цели кредита, среднемесячный доход и расход и другие. Расхождения в характеристиках третьего сегмента меньше, чем в трёх других сегментах.

Результаты табл. 3 также соответствуют построенным матрицам расстояний, приведённым на рис. 9, то есть плавные переходы цветов от одного класса к другому соответствуют общим характеристикам для данных сегментов.

Как показали результаты проведенных исследований, диаграмма Хинтона хорошо отображает разделение исходного множества на известное число сегментов, в данной ситуации – 4, что подтверждают результаты, приведенные в табл. 3. Карты Кохонена как аппарат интеллектуального анализа могут применяться в банковских системах для сегментации клиентов и многих других областях человеческой деятельности, где возникают задачи кластеризации и классификации, и необходима визуализация более чем двумерного входного пространства.

Выводы

В работе рассмотрены различные способы визуализации нейросетевых карт Кохонена для выявления скрытых структур в данных низкой и высокой размерности. Проведены экспериментальные исследования построения карт Кохонена для двумерных данных на примере набора данных с координатами двух спиралей. Применены способы визуализации посредством унифицированной матрицы расстояний и визуализация диаграммой Хинтона.

Экспериментальные исследования, проведённые для оценки качества визуализации, с помощью матрицы расстояний показали, что более эффективным является построение карты большего размера, чем добавление ячеек для отображения расстояния между нейронами при раскраске карты. Полученные результаты позволили сделать заключение о том, что при увеличении размера топологической карты свободные ячейки заполняются неактивными нейронами, которые фактически заполняют ячейки, визуализирующие расстояние, или то же отображают активные нейроны в зависимости от размера карты и распределения образов по узлам.

Также проведен эксперимент по построению и визуализации карт Кохонена для многомерного набора данных из области банковского кредитования. Данные о клиентах были разделены сетью на 4 группы и отображены посредством диаграммы Хинтона.

Полученные результаты свидетельствуют об эффективности применения аппарата нейронных сетей Кохонена для визуализации данных различной размерности, с различными структурами.

Перспективы и дальнейшие исследования направлены на применение средств визуализации к другим нейронным сетям, решающим аналогичные задачи. Будет рассмотрена возможность применения данной нейронной сети в качестве модуля аналитической системы для принятия решений при кредитовании в банке, а также разработаны другие карты для визуализации работы нейронной сети Кохонена.

Список литературы: 1. *Зиновьев А.Ю.* Визуализация многомерных данных. — Красноярск: Изд-во КГТУ, 2000. — 168 с. 2. *Бодянский Е.В., Руденко О.Г.* Искусственные нейронные сети: Уч. пособие. – Харьков: ООО «Компания СМИТ», 2005. – 408 с. 3. *Калан Р.* Основные концепции нейронных сетей.: Пер. с англ. – М.: Изд. дом «Вильямс», 2003. – 288 с. 4. *Хайкин С.* Нейронные сети: полный курс, 2-е изд. – М.: ООО «И.Д. Вильямс», 2006. – 1104 с. 5. *Kohonen T.* Self-organizing maps / Teuvo Kohonen. – 3. ed. – Berlin: Springer-Verlag Berlin Heidelberg, 2001. – 500 p.

Поступила в редколлегию 18.04.2008