DOI: 10.5862/JPM.242.11
УДК 004.032.26

А.А. Пастухов, А.А. Прокофьев

Национальный исследовательский университет
«Московский институт электронной техники»

ПРИМЕНЕНИЕ САМООРГАНИЗУЮЩИХСЯ КАРТ КОХОНЕНА ДЛЯ ФОРМИРОВАНИЯ ПРЕДСТАВИТЕЛЬСКОЙ ВЫБОРКИ ПРИ ОБУЧЕНИИ МНОГОСЛОЙНОГО ПЕРСЕПТРОНА

Рассмотрен вопрос эффективного формирования представительской выборки для обучения нейронной сети типа многослойный персептрон. Обозначены основные проблемы, возникающие в процессе разбиения факторного пространства на тестовое, проверочное и обучающее множества. Предложен подход, основанный на применении кластеризации, позволяющий увеличить энтропию обучающего множества. Рассмотрены самоорганизующиеся карты Кохонена как эффективный метод кластеризации. На базе таких карт проведена кластеризация факторных пространств различной размерности и сформирована представительская выборка. Синтезирована и обучена нейронная сеть типа многослойный персептрон на множестве, сформированном с использованием и без использования кластеризации. Сделан вывод о том, что рассматриваемый подход способствует повышению энтропии обучающего множества и, как следствие, приводит к улучшению качества обучения многослойного персептрона при небольшой размерности факторного пространства.

ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ, САМООРГАНИЗУЮЩАЯСЯ КАРТА КОХОНЕНА, КЛАСТЕРИЗАЦИЯ, ПРЕДСТАВИТЕЛЬСКАЯ ВЫБОРКА.

Введение

Обучение нейронной сети – это важный этап ее функционирования. Для обучения многослойного персептрона (MLP – multilayer perceptron) наиболее часто используют алгоритм обратного распространения ошибки.

Перед процедурой обучения MLP особое внимание уделяется предобработке данных. В большинстве работ по применению нейронных сетей методика предобработки сводится к нормализации, масштабированию, а также начальной инициализации весов.

Данные действия, несомненно, необходимы, но их нельзя считать достаточными. При небольшой размерности факторного пространства следует учитывать специфику распределения исходных данных для эффективного обучения нейронной сети. При большом же количестве факторов это слишком сложно сделать. В последнем случае целесообразно применять кластеризацию для формирования обучающего множества из примеров признаков, наиболее уникальных по совокупности.

Один из эффективных методов кластеризации – использование самоорганизующихся карт Кохонена. Они нашли широкое применение в различных областях. Так, например, в работах [1, 2] описано применение указанных карт для распознавания образов. Кроме того, они эффективны при создании систем тестирования [3], анализа состава растворов [4]; на их основе разрабатываются новые модели для кластеризации данных [5]. Описание архитектуры, процедуры обучения и примеры использования самоорганизующихся карт Кохонена представлено, например, в работах [6, 7].

Как отмечено выше, проведение кластеризации факторного пространства позволяет формировать представительскую выборку, содержащую наиболее уникальные по совокупности признаков обучающие примеры, для обучения многослойного персеп-

трона. Аналогичный подход, включающий использование самоорганизующихся карт Кохонена для кластеризации, встречается, например, в работе [8].

В нашей работе исследовано применение кластеризации на основе самоорганизующихся карт Кохонена, но с точки зрения увеличения энтропии обучающего множества; проанализированы также эффективность такого подхода для факторных пространств различных размерностей и влияние размерности на изменение энтропии обучающего множества при использовании кластеризации.

При обучении нейронной сети типа MLP по алгоритму обратного распространения ошибки ответственным является этап формирования факторного пространства, на которое налагаются следующие требования:

1) необходима непротиворечивость данных, участвующих в обучении;

2) должны присутствовать максимально уникальные по совокупности признаки примеров, составляющие обучающее множество;

3) необходимо достаточное количество обучающих данных для сети выбранной архитектуры.

Чтобы отвечать первому требованию, обучающее множество должно быть проанализировано на наличие противоречий, необходимо выяснить причины возникновения ошибок (ошибка появилась при внесении данных или, что более серьезно, в результате использования недостаточного количества признаков факторного пространства) и по возможности их устранить.

Удовлетворять второму требованию необходимо для того, чтобы максимально эффективно использовать обучающую выборку. Количество данных, используемых для обучения нейронной сети, часто бывает невелико, поэтому крайне важно правильно сформировать обучающее множество, содержащее данные, наиболее уникальные по совокупности признаков.

Третье требование предъявляется для того, чтобы достичь заданной точности обучения нейронной сети за конечное количество шагов. В работе [9] приведена зависимость ошибки обучения от количества свободных параметров W (архитектуры нейронной сети) и количества обучающих примеров N:

$$N = O(W / \varepsilon), \qquad (1)$$

где ε – допустимая точность ошибки обучения; $O(...)$ – порядок величины, заключенной в скобки.

Нами были исследованы способы формирования обучающего множества, содержащего наиболее уникальные по совокупности признаков примеры, за счет увеличения энтропии. В данной статье под термином энтропия подразумевается неопределенность выбора примера из обучающего множества.

Для повышения вероятности адекватного обучения многослойного персептрона факторное пространство разбивается на три множества: обучающее, тестовое и проверочное [10]. Первое используется для настройки свободных параметров нейронной сети, второе – для независимого тестирования уже обученной нейронной сети, третье – для исключения эффекта переобучения, который заключается в запоминании, а не обобщении обучающего множества.

NNtool Box пакета MatLab использует для обучения 80 % случайно выбранных векторов из факторного пространства. Такое разбиение нельзя считать оптимальным, поскольку крайне мала вероятность выбора векторов, уникальных по совокупности признаков, т. е. такого разбиения, при котором энтропия обучающего множества максимальна и равна $\log_2 N_t$ (N_t – размер обучающего множества).

Таким образом, следует считать актуальной разработку метода, который бы позволял достигать максимума энтропии обучающего множества (если это позволяет характер данных, составляющих факторное пространство) либо гарантированно достигать определенного значения энтропии указанного множества (большего, чем при случайном разбиении факторного пространства на представительскую выборку).

С целью повышения энтропии обучающего множества предлагается провести кластерный анализ [11] факторного про-

странства с тем, чтобы разбить последний на обучающее, тестовое и проверочное подмножества для формирования представительской выборки.

Чтобы эффективно применять алгоритмы кластерного анализа, очень важно правильно определить число прототипов. Одним из надежных способов кластеризации следует считать метод, основанный на применении самоорганизующихся карт Кохонена [6]. Чтобы провести кластеризацию с использованием таких карт, требуется также указать число прототипов, однако благодаря самоорганизации и обучению без учителя, сеть способна самостоятельно определять центры кластеров. Кроме того, следует отметить простоту реализации самоорганизующихся карт Кохонена, а также гарантированное получение ответа после прохождения данных по слоям карты.

Таким образом, представляется целесообразным использовать самоорганизующиеся карты Кохонена для кластеризации факторного пространства, а затем анализировать результаты обучения многослойного персептрона на представительской выборке, полученные с применением и без применения предлагаемого подхода.

Для проведения эксперимента были сгенерированы исходные данные, которые формируют факторное пространство. Последнее определяется пятью параметрами: из них четыре входных (x_1, x_2, x_3, x_4) и один выходной (y). Связь между этими параметрами задана функцией

$$y = e^{x_1} + e^{x_2} + 2e^{x_3} + 3e^{x_4}. \qquad (2)$$

Кроме того, во входной вектор добавлен шум, который описан случайной величиной, распределенной по нормальному закону с дисперсией 0,01. Эксперимент был проведен на десяти факторных пространствах, включающих от 100 до 1000 векторов.

Постановка задачи

Пусть

$$X = \{X^1, ..., X^M, Y^1, ..., Y^M\}$$

– факторное пространство,
где $X^i = \{x_1, x_2, x_3, x_4\}$, $Y^i = \{y(X^i)\}$, M – количество векторов в факторном пространстве.

Требуется найти с применением самоорганизующихся карт Кохонена такое разбиение факторного пространства на три множества (T – обучающее, V – проверочное и E – тестовое), для которого выполняется условие

$$H_0(T) < H(T) \le H_{\max}(T), \qquad (3)$$

где $H(T)$, $H_0(T)$ – величины энтропии обучающего множества с использованием кластеризации и для случайного разбиения факторного пространства на представительскую выборку, соответственно; $H_{\max}(T) = \log_2 N_t$ – максимальная энтропия этого множества (N_t – размер обучающего множества, составляющего 80 % от факторного пространства).

Описание нейронной сети Кохонена

Указанная нейронная сеть, или самоорганизующаяся карта признаков, имеет набор входных элементов, число которых совпадает с размерностью векторов, составляющих факторное пространство, и имеет набор выходных элементов, соответствующих кластерам (кластерные элементы – КЭ).

Входные элементы предназначены для распределения входного вектора между выходными элементами сети. Весовые значения КЭ можно интерпретировать как значения координат, описывающих позицию кластера в пространстве входных данных.

В работе [9] отмечается, что КЭ целесообразно располагать в форме двумерной решетки, так как такая топология гарантирует, что каждый нейрон будет иметь множество соседей. От этого расположения зависит, какие элементы будут корректироваться в радиусе кластерного элемента-победителя. Множество корректируемых КЭ определяется нормой, выбранной в пространстве весов; этой норме соответствует геометрия окрестности выбранного радиуса. В простейшем случае КЭ равен единице (корректируются веса только элемента-победителя).

На рис. 1 распределительный слой (DL) соответствует входному, а слой Кохонена

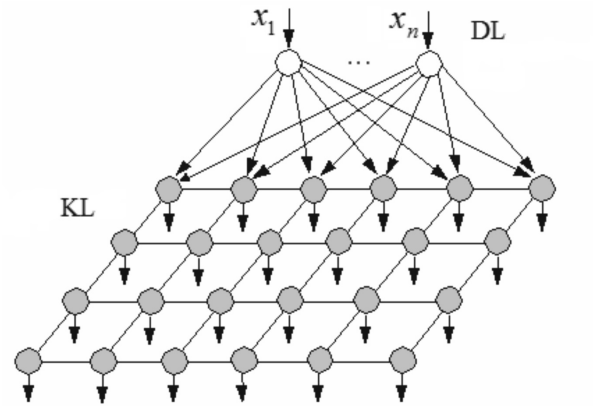

Рис. 1. Нейронная сеть Кохонена:
$x_1, ..., x_n$ – входные параметры факторного пространства; DL, KL – распределительный слой и слой Кохонена соответственно; кластерные элементы изображены шариками

(KL) содержит КЭ, образующие прямоугольник.

Обучение сети Кохонена проходит в два этапа [9]. На первом вычисляются расстояния от обучающих образцов до каждого КЭ (нейрона) по формуле [12]:

$$d_j = \sum_i (\omega_{ij} - x_i)^2,$$

где ω_{ij} – весовой коэффициент, связывающий входной вектор x_i с кластерным элементом j.

Значения d_j передаются в конкурирующую функцию активации передачи, которая возвращает нулевое значение для всех выходных нейронов кроме нейрона-победителя k. Нейроном-победителем считается нейрон, для которого выполняется условие

$$d = \min(d_j),$$

где $1 \le j \le n$ (n – количество КЭ).

Весовой вектор нейрона-победителя расположен ближе всех к входному вектору и поэтому его выход устанавливается равным единице.

На втором этапе производится корректировка весовых коэффициентов нейрона-победителя k и всех нейронов из заданного радиуса r, т. е. уточняются позиции КЭ в пространстве входных данных. Для корректировки весовых коэффициентов мы применяли формулу

$$\omega_{ij}(n+1) = \omega_{ij}(n) + \eta(n)[x_i - \omega_{ij}(n)],$$

где $\omega_{ij}(n)$ – весовой коэффициент, связывающий входной вектор x_i с кластерным элементом j на итерации n; $\eta(n) \in [0;1]$ – коэффициент скорости обучения, позволяющий управлять величиной коррекции весовых коэффициентов на каждой итерации.

Можно применять и другие способы корректировки весовых коэффициентов, например, с использованием функции окрестности [9], но данный вопрос выходит за рамки нашего исследования.

Коэффициент скорости обучения обычно инициализируется достаточно большой величиной (близка к единице), которая по мере обучения уменьшается.

Радиус r также изначально инициализируется достаточно большим значением и уменьшается на каждой итерации вплоть до одного элемента-победителя. Закон изменения радиуса подбирается экспериментально. В простейшем случае радиус на каждом шаге уменьшается по линейному закону.

Критерием окончания процесса обучения служит значение величины изменения весовых коэффициентов на очередной итерации: если она меньше заданного значения, то процесс завершен.

Кластерный анализ факторного пространства с применением нейронной сети Кохонена

Как упоминалось выше, было сформировано факторное пространство с пятью параметрами для обучения многослойного персептрона (MLP) моделированию целевой функции (2). NNToolbox пакета MatLab разбивает это пространство на три множества случайным образом в пропорции 8 : 1 : 1 [13].

Далее в нашей статье приведены расчеты для факторного пространства размера $N = 100$. Расчеты для пространств большего размера ($N \in [200, ..., 1000]$) проводились аналогичным образом, поэтому опущены. В конце раздела приведена лишь сводная таблица результатов расчетов для $N \in [100, ..., 1000]$.

Случай $N = 100$. Представительская выборка формируется NNToolbox пакета MatLab следующим образом: 80 : 10 : 10 векторов, где значения отвечают соответственно обучающему, проверочному и тестовому множествам.

Восемьдесят векторов, формирующих тестовое множество, в данном случае не учитывают специфику входных признаков, и поэтому такой выбор обучающих примеров нельзя считать оптимальным в отношении максимизации энтропии.

Ввиду вышеизложенного, предлагается выбирать для обучающего множества на основе кластеризации, проведенной сетью Кохонена, примеры из кластеров с минимальным количеством элементов (в идеальном случае из кластеров, представленных одним элементом). Суть предложения заключается в том, что в идеальном эксперименте обучающее множество должно содержать по одному элементу из каждого кластера. В случае если нейронная сеть Кохонена определила малое количество кластеров, в обучающее множество добавляются еще по одному элементу из каждого кластера. Однако необходимо стремиться к тому, чтобы количество представителей из неединичных кластеров было обратно пропорционально размеру кластера.

С учетом этого была выбрана прямоугольная топология сети Кохонена размерностью 8 × 10. Конфигурации прямоугольника подбирается экспериментально (возможны варианты 10 × 8 или 4 × 20). В данном случае при конфигурации 8 × 10 количество единичных кластеров оказалось максимальным, а в каждом неединичном кластере количество примеров оказалось минимальным.

Инициализация карты Кохонена, т. е. присвоение начальных значений весам нейронов, может быть проведена разными способами. В данной работе карта инициализирована малыми случайными значениями, но существуют и более продвинутые алгоритмы начальной инициализации [14].

В нашем случае сеть Кохонена разбила факторное пространство на 73 кластера (рис. 2), представляющее собой прямоугольную таблицу размером 8 × 10. Видно,

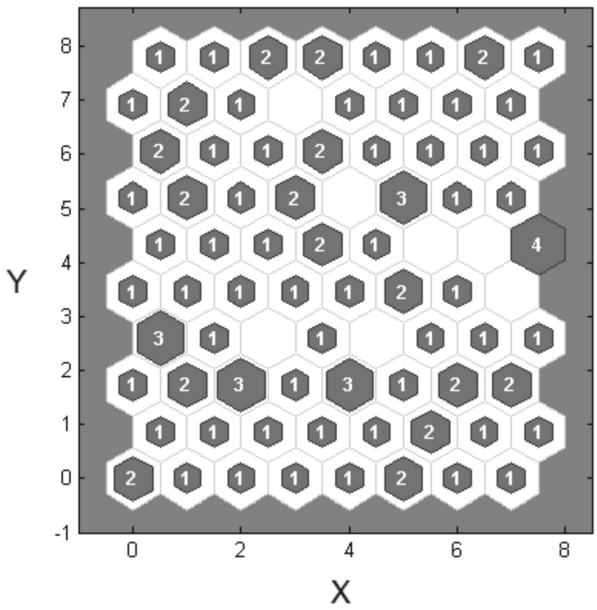

Рис. 2. Распределение факторного пространства по 80 кластерам, построенное сетью Кохонена (прямоугольная таблица размером 8 × 10).
Числа в ячейках отвечают количеству элементов в кластерах. Пара (X,Y) однозначно определяет положение кластера в пространстве карты Кохонена

что семь кластеров не содержат ни одного элемента, что, тем не менее, не оказывается критичным. В этой связи для повышения энтропии обучающего множества факторное пространство было разбито на представительскую выборку следующим образом: 73 : 10 : 17 наборов (обучающее/проверочное/тестовое множества), причем для обучающего множества было взято по одному представителю из кластера.

С учетом этого разбиения факторного пространства вычислялась энтропия обучающего множества для его случайного разбиения на представительскую выборку (в этом случае наборы для обучения выбираются случайно из всего факторного пространства) и для разбиения с учетом кластеризации (наборы для обучения выбираются из имеющихся 73-х, определенных с помощью кластеризации).

Для случая $N = 100$ существуют кластеры, содержащие в себе соответственно 1, 2, 3 и 4 элемента (см. рис. 2). Обозначим их $K = \{K_1, K_2, K_3, K_4\}$.

Таблица 1

Классификация кластеров по количеству элементов

$K_i \in K$	K_1	K_2	K_3	K_4
$\sum K_i$	52	16	4	1

В табл. 1 приведено соотношение количеств кластеров, относящихся к соответствующему подмножеству множества K.

Для текущего разбиения на кластеры вычислим энтропию для факторного пространства так, если бы формирование представительской выборки осуществлялось случайным образом.

Рассмотрим факторное пространство в качестве системы из 73 элементов (на столько кластеров разбила факторное пространство сеть Кохонена). Если формировать обучающее множество случайным образом, то вероятность p_i выбора элемента из кластера K_i будет равна i / 100 (i – количество элементов в кластере).

Энтропия для дискретных случайных событий вычисляется по формуле Шенона [15]:

$$H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i. \qquad (4)$$

Используя эту формулу, получим искомое значение:

$$H(x) = -\left(\sum_{i=1}^{52} \frac{1}{100} \log_2 \frac{1}{100} + \sum_{i=1}^{16} \frac{2}{100} \log_2 \frac{2}{100} + \right.$$

$$\left. + \sum_{i=1}^{4} \frac{3}{100} \log_2 \frac{3}{100} + \frac{4}{100} \log_2 \frac{4}{100} \right) \approx 6,05 \text{ бит}.$$

В случае формирования обучающего множества из представителей каждого кластера (по одному), получим равновероятные события выбора, т. е. вероятность такого события равна 1/73. С помощью формулы (4) получим энтропию, равную $\log_2 n = 6,19$, где $n = 73$. Таким образом, прирост энтропии тестового множества составляет 0,14 бит.

Максимально возможная энтропия, равная 6,32 ($\log_2 80 = 6,32$) для случая $N = 100$, гипотетически достигается в случае, если все 80 примеров являются абсолютно уникальными по совокупности признаков.

Таблица 2

Результаты расчета энтропии для факторных пространств различной размерности

N	Энтропия, бит			Затраченное время, с	
	$H_{\max}(T)$	$H(T)$	$H_0(T)$	T_1	T_2
100	6,32	6,19	6,05	3	1
200	7,32	7,15	7,02	5	2
300	7,91	7,79	7,67	7	4
400	8,32	8,17	8,07	10	3
500	8,64	8,50	8,38	15	3
600	8,91	8,77	8,66	21	4
700	9,13	8,97	8,86	29	9
800	9,32	9,17	9,05	36	7
900	9,49	9,31	9,16	42	8
1000	9,64	9,45	9,29	53	11

Обозначения: N – количество элементов факторного пространства; $H_{\max}(T)$ – максимальная энтропия обучающего множества; $H(T)$, $H_0(T)$ – величины энтропии этого множества с использованием кластеризации и для случайного разбиения факторного пространства на представительскую выборку, соответственно; T_1, T_2 – промежутки времени, затраченные соответственно на обучение самоорганизующейся карты Кохонена (размер $0,8N$) и многослойного персептрона на данных соответствующей размерности (архитектура многослойного персептрона выбиралась в соответствии с формулой (4) из расчета $\varepsilon = 0,2$.

Результаты аналогичных расчетов для факторных пространств с количеством векторов от 100 до 1000 приведены в табл. 2.

Анализ данных табл. 2 приводит к выводу, что для любого N выполняется условие (3). Значение энтропии для случая с использованием кластеризации лежит примерно посередине между значениями $H_0(T)$ и $H_{\max}(T)$ для всех N. Временные затраты на обучение карты Кохонена растут практически линейно.

Следует отметить, что время обучения многослойного персептрона (T_2) растет медленнее, чем время обучения самоорганизующейся карты Кохонена (T_1). Возможно, анализ времени обучения на данных более сложных количественной и качественной структур был бы показательнее. Этот вопрос требует дальнейшего исследования.

Можно констатировать, что для факторного пространства очень большой размерности временные затраты могут быть неприемлемы, но для небольших факторных пространств использование кластеризации гарантирует сокращение размера обучающего множества и в то же время прирост его энтропии.

На следующем этапе исследования ставился эксперимент, состоящий в обучении многослойного персептрона с использованием как представительской выборки, сформированной случайным образом, так и сформированной на основе кластеризации.

Обучение многослойного персептрона с использованием кластеризации и без нее

Нами проведены две процедуры обучения нейронной сети типа MLP на тестовых данных по методу обратного распространения ошибки [16]. В первом случае использовали формирование представительской выборки, предложенное NNToolBox пакета MATLAB, во втором применяли подход на основе кластерного анализа. В эксперименте использовалось факторное пространство, включающее 100 элементов.

Для контрольного сравнения с результатами исследования была также проведена процедура обучения по алгоритму, отличному от стандартного обратного распространения ошибки, в частности, по квазиньютоновскому методу обратного распространения (BFGS).

Нейронная сеть типа MLP в обоих случаях имеет одинаковую архитектуру: 4-4-1. Результаты обучения сети для обоих случаев представлены на рис. 3 – 7.

Графики регрессии (рис. 3 и 4) для обоих случаев аналогичны по своему виду. Табл. 3, 4 содержат результаты анализа данных, приведенных на рис. 3 и 4, соответственно. Можно констатировать, что отклонение от тренда составило в обоих случаях намного меньше 1 %.

Анализ результатов, представленных на рис. 5, позволяет сделать следующие выводы:

1. Наилучшая производительность нейронной сети (минимальная величина среднеквадратичной ошибки (MSE)) для проверочного множества при формировании представительской выборки без кластеризации составляет 0,31462, а в случае использования кластеризации – 0,11601.

2. Разница в среднеквадратичной ошибке между обучающим и тренировочным/проверочным множествами оказывается существенно выше в случае, когда кластеризация не используется.

3. Наилучшая производительность (минимальная величина среднеквадратичной ошибки) без использования кластеризации была достигнута на 165-й эпохе обучения против 254-й с использованием кластеризации, однако, как отмечалось выше, для первого случая наилучшая производительность (на рис. 5 минимальное значение на графике Validation) составила 0,31462, а с использованием кластеризации – 0,11601.

Сопоставление результатов, представленных на рис. 6 и 7, позволяет заключить, что значение градиента поверхности отклика ошибки на рис. 6 по окончании процедуры обучения оказывается на порядок ниже, чем на рис. 7, где представлено разбиение на представительскую выборку случайным образом. Поверхность отклика ошибки обучения есть пространство размерности n, где n – количество входных параметров факторного пространства (было взято $n = 4$). Вектор градиента поверхности

ошибок указывает направление кратчайшего спуска по этой поверхности из данной точки к минимуму (но не обязательно глобальному) ошибки обучения. Значение градиента для функции поверхности отклика $u(x_1, x_2, ..., x_n)$ определяется через частные производные:

$$g = \sqrt{\left(\frac{\partial u}{\partial x_1}\right)^2 + ... + \left(\frac{\partial u}{\partial x_n}\right)^2}.$$

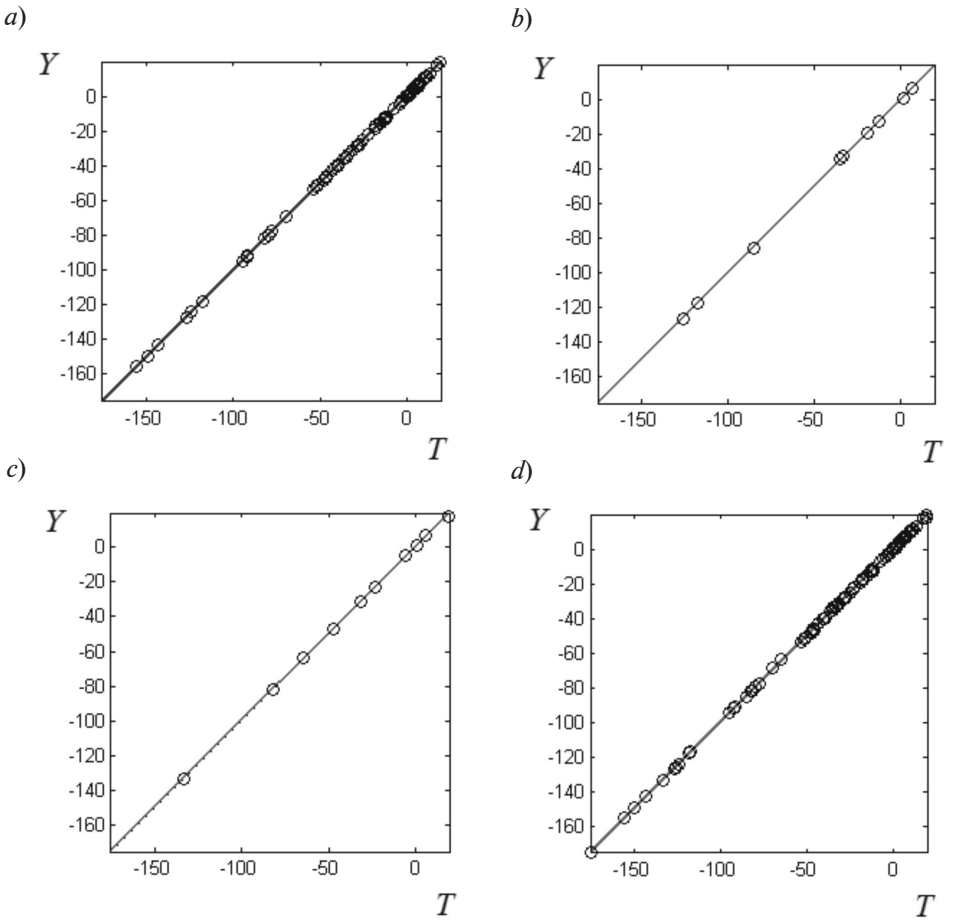

Рис. 3. Графики регрессии обучения нейронной сети со случайным формированием обучающего множества:
Результаты обучения (*a*), проверки на переобучение (*b*) и проверки на тестовом множестве (*c*), а также общий результат (*d*). Прямая на графике проходит через центр облака данных

Таблица 3

Результаты обучения нейронной сети со случайным формированием обучающего множества

График на рис. 3	*R*	*Y*(*T*)
a	0,99998	*T* + 0,00150
b	0,99993	*T* + 0,00046
c	0,99994	*T* + 0,01900
d	0,99997	*T* + 0,00240

Обозначения: *R* – показатель отношения «значение выхода нейронной сети / целевое значение», *Y*(*T*) – приближенная линейная зависимость фактических величин значения функции от целевых величин *T*.

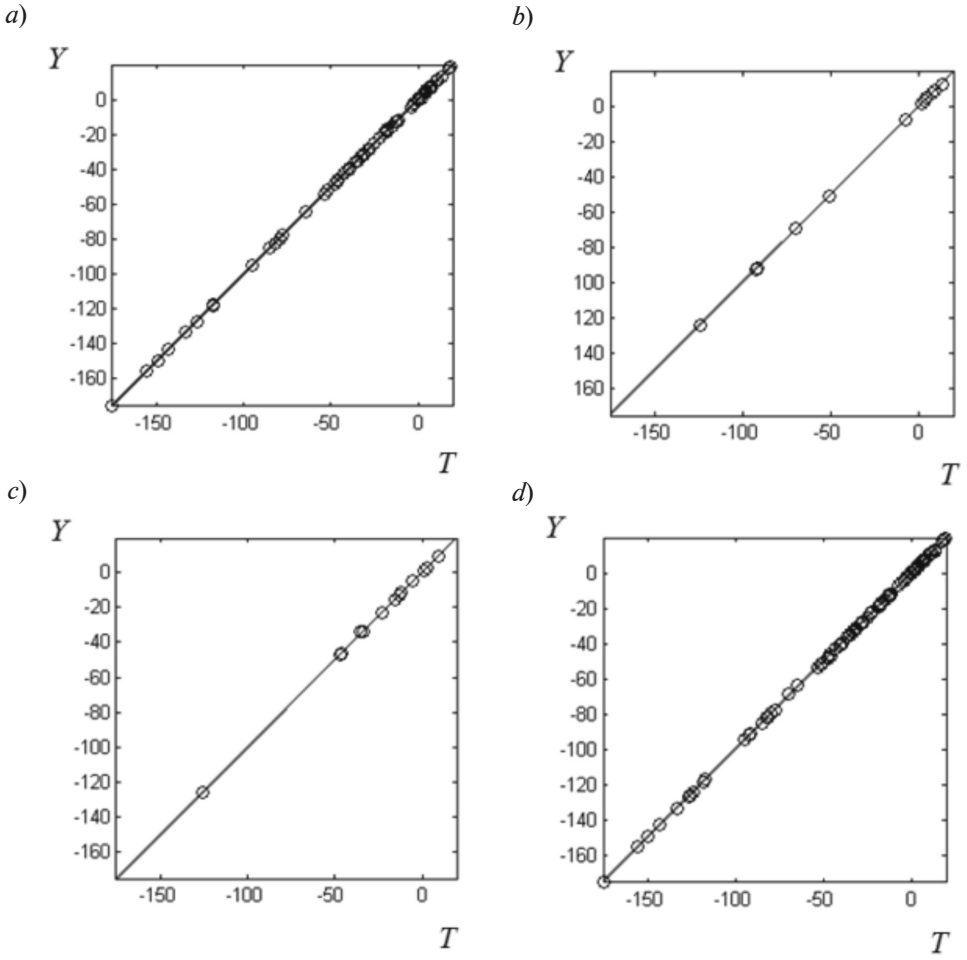

Рис. 4. Графики регрессии обучения нейронной сети с формированием обучающего множества (использована кластеризация)
(обозначение величин см. в подписи к рис. 3)

Таблица 4

Результаты обучения нейронной сети с формированием обучающего множества (использована кластеризация)

График на рис. 4	*R*	*Y*(*T*)
a	0,99998	*T* + 0,0016
b	0,99998	*T* + 0,0600
c	0,99994	*T* + 0,1700
d	0,99997	*T* + 0,0280

Обозначение величин дано в подписи к табл. 3.

Следует отдельно отметить, что малое значение величины градиента не служит показателем повышения качества обучения в том случае, если процесс такого обучения сошелся к локальному минимуму, который отличается от глобального.

Использование квазиньютоновского метода обратного распространения (BFGS) для пространств с количеством элементов, близким к 1000, дало лучшие результаты, чем применение стандартного алгоритма обратного распространения, однако для

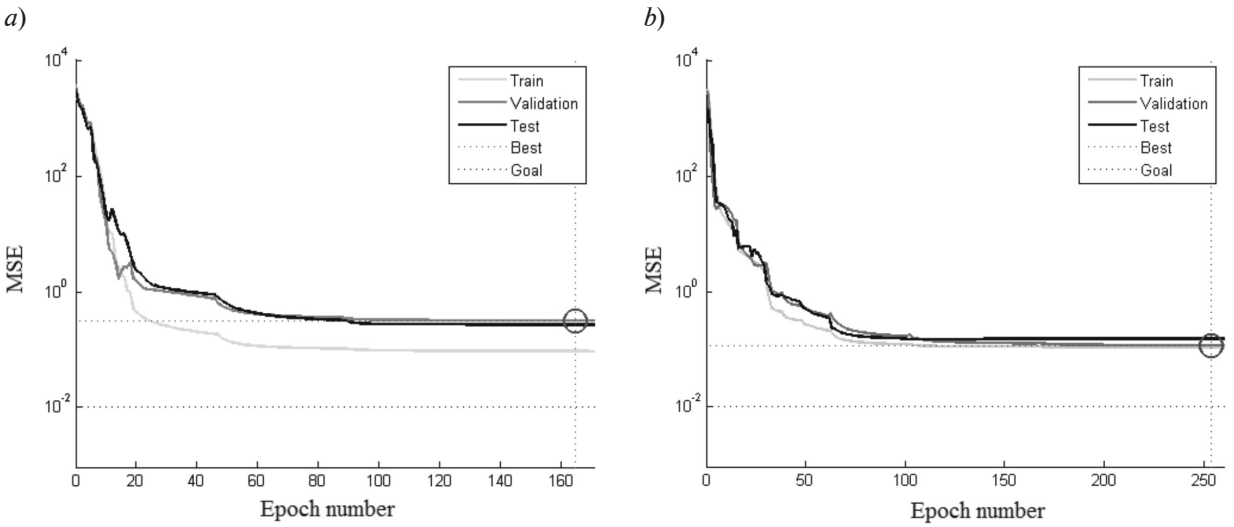

Рис. 5. Графики производительности нейронной сети в процессе обучения со случайным формированием обучающего множества (*a*) и с применением кластеризации (*b*):
MSE – среднеквадратичная ошибка обучения; Epoch number – его текущая эпоха; представлено поведение ошибки для обучающего (Train), проверочного (Validation) и тестового (Test) множеств; Goal, Best – целевое и наилучшее значения ошибки, последнее достигнуто для проверочного множества μ

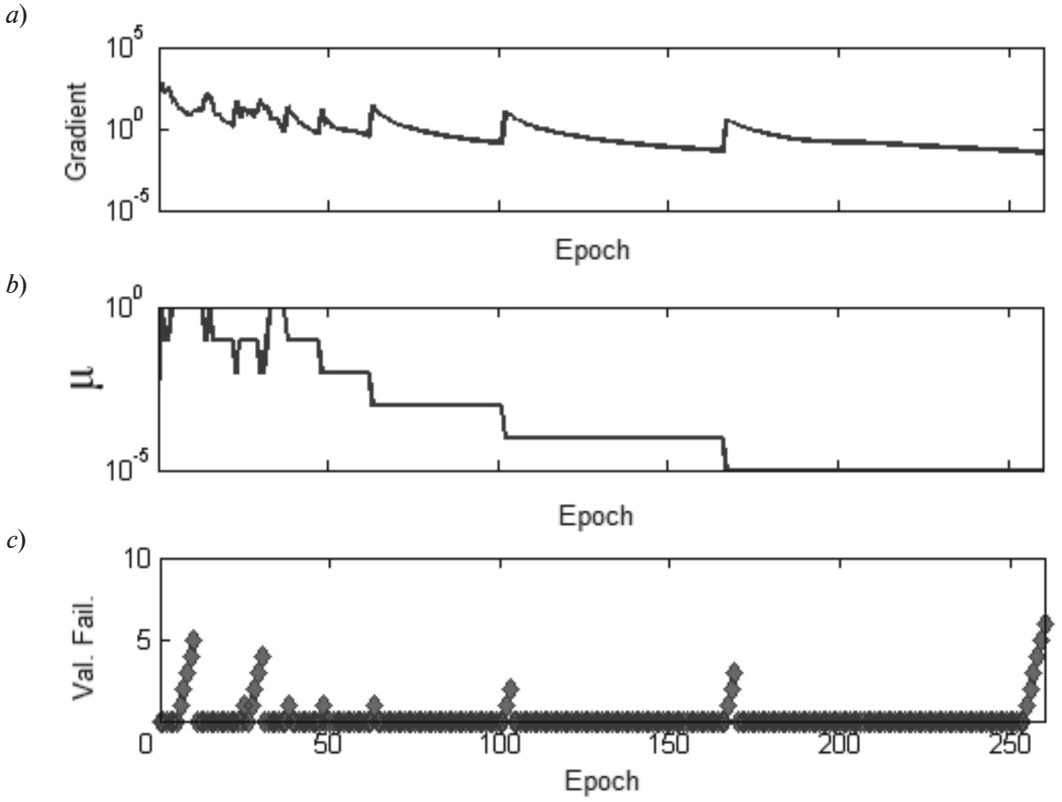

Рис. 6. Графики состояния обучения со случайным формированием тестового множества. Представлены зависимости градиента (Gradient) (*a*), адаптации μ (*b*) и количества проверок на переобучение (Validation Fail) (*c*) на соответствующих значениях эпохи (Epoch).
Значение градиента по окончании процедуры обучения оказалось равным 0,033976, адаптации – 10^{-5}; количество проверок на переобучение – 6

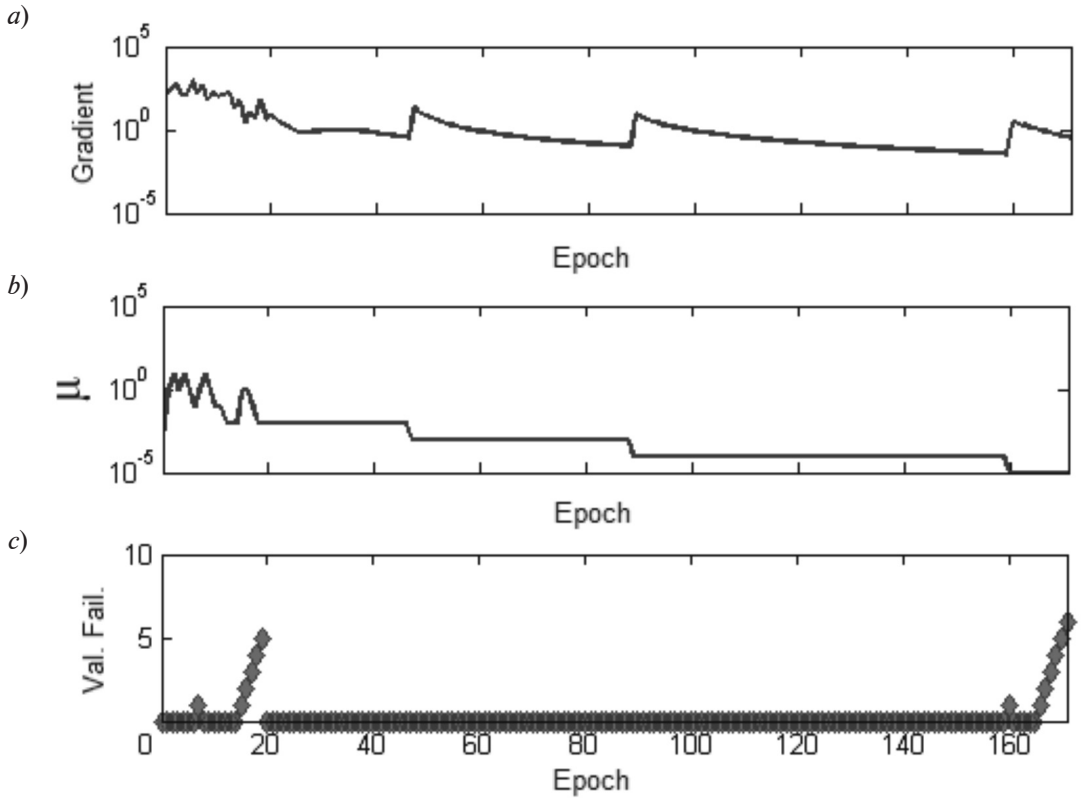

Рис. 7. Графики состояния обучения с применением кластеризации (обозначения величин см. в подписи к рис. 6).
В данном случае значение градиента по окончании процедуры обучения оказалось равным 0,35385, адаптации – 10^{-5}; количество проверок на переобучение – 6

пространств малой размерности (например, для $N = 100$) применение стандартного алгоритма обратного распространения ошибки на основе наискорейшего спуска оказалось эффективней в отношении конечного значения среднеквадратичной ошибки обучения.

Заключение

Подведение итогов для полученных результатов позволяет заключить, что предложенный нами подход позволяет успешно решить поставленную задачу. Энтропия обучающего множества при использовании кластеризации для формирования обучающего множества увеличилась и приблизилась к максимально возможному значению.

Несмотря на то, что при использовании кластеризации обучающее множество включает меньшее количество примеров, чем при случайном разбиении его на представительскую выборку, разница между ошибками для обучающего и тестового/проверочного множеств очевидно меньше, когда применяется кластеризация. Полученный результат наглядно показывает повышение качества обучения. Кроме того, величина среднеквадратичной ошибки оказывается значительно меньше в случае разбиения на представительскую выборку с применением кластеризации.

СПИСОК ЛИТЕРАТУРЫ

[1] **Тант Зин Пьо, Тин Чжо, Пья Сон Ко Ко, Пайе Тэйн Наинга.** Методика системы распознавания образов с помощью самоорганизующихся карт Кохонена нейронных сетей на основе Matlab. Интернет-журнал «Науковедение». 2013. № 5 [http://naukovedenie.ru/PDF/27tvn513.pdf].

[2] **Kumar D., Rai C.S., Kumar S.** Face recognition using self-organizing map and princi-

pal component analysis // Proc. on Neural Networks and Brain (ICNNB 2005). Oct. 2005. Vol. 3. Pp. 1469–1473.

[3] **Панфилова А.С.** Система тестирования интеллекта на базе факторных моделей и самоорганизующихся карт Кохонена // Нейрокомпьютеры: разработка, применение. 2012. № 9. С. 6 –12.

[4] **Гущин К.А., Доленко С.А., Буриков С.А., Доленко Т.А.** Применение алгоритмов кластеризации и понижения размерности данных в задачах анализа состава многокомпонентных растворов // XIII Всерос. научн. конф. «Нейрокомпьютеры и их применение». Тез. докл. М.: МГППУ, 2015. С. 72–73.

[5] **Новиков А.В.** Нейросетевые методы решения задач кластерного анализа. // Нейрокомпьютеры: разработка, применение. 2014. № 2. С. 48 –53.

[6] **Кохонен Т.** Самоорганизующиеся карты. М.: Бином. Лаборатория знаний, 2008. 655 с.

[7] **Горбаченко В.И.** Нейроинформатика. Конспект лекций. Пенза: Пензенский государственный педагогический университет, 2011. 81 с.

[8] **Ковалев И.В.** Интеллектуальная система прогнозирования загрязнения атмосферы // Нейрокомпьютеры: разработка, применение. 2010. № 7. С. 62–66.

[9] **Хайкин С.** Нейронные сети: полный курс. 2-е изд. Пер. с англ. М.: ИД «Вильямс», 2008. 1104 с.

[10] **Бэстенс Д.Э., Ван Ден Берг В.М., Вуд Д.** Нейронные сети и финансовые рынки: принятие решений в торговых операциях. М.: ТВП,1997. 236 с.

[11] **Паклин Н.Б., Орешков В.И.** Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2013. 704 с.

[12] **Калан Р.** Основные концепции нейронных сетей. М.: Вильямс, 2001. 288 с.

[13] **Beale M.H., Hagan M.T., Demuth H.B.** Neural Network Toolbox User's Guide [http://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf] MathWorks, Inc. 2014. 410 p.

[14] **Горбачев В.И.** Сети и карты Кохонена. URL: http://gorbachenko.self-organization.ru/articles/Self-organizing_map.pdf (дата обращения: 11.01. 2016).

[15] **Дмитриев В.И.** Прикладная теория информации. Учеб. пос. для студ. вузов. М.: Высшая школа, 1989.

[16] **Пастухов А.А.** Применение нейронных сетей для прогнозирования параметров энергетических установок с лазерным зажиганием // Научно-технические ведомости СПбГПУ. Физико-математические науки. 2015. № 2 (218) С. 19–29.

СВЕДЕНИЯ ОБ АВТОРАХ

ПАСТУХОВ Алексей Андреевич – *аспирант кафедры высшей математики № 1 Национального исследовательского университета «МИЭТ».*

124498, Российская Федерация, Москва, Зеленоград, проезд 4806, д. 5

pastuhov1992@gmail.com

ПРОКОФЬЕВ Александр Александрович – *доктор педагогических наук, заведующий кафедрой высшей математики № 1 Национального исследовательского университета «МИЭТ».*

124498, Российская Федерация, Москва, Зеленоград, проезд 4806, д. 5

aaprokof@yandex.ru

Pastukhov A.A., Prokofiev A.A. KOHONEN SELF-ORGANIZING MAP APPLICATION TO REPRESENTATIVE SAMPLE FORMATION IN THE TRAINING OF THE MULTILAYER PERCEPTRON.

In this paper, we have considered an item of effective formation of a representative sample for training the neural network of the multilayer perceptron (MLP) type. The main problems arising in the process of the factor space division into the test, verification and training sets were formulated. An approach based on the use of clustering, that allowed one to increase the entropy of the training set was put forward. Kohonen self-organizing maps (SOM) were examined as an effective procedure of a clustering. Based on such maps, the clustering of factor spaces of different dimensions was carried out, and a representative sample was formed. To verify our approach we synthesized the MLP neural network and trained it. The training technique was performed with the sets formed both using the clustering and no doing it. The approach under consideration was concluded to have an influence on the increase in the entropy of the training set and (as a result) to lead to the quality improvement of training of MLP with the small dimensionality of the factor space.

ARTIFICIAL NEURAL NETWORK, KOHONEN SELF-ORGANIZING MAP, CLUSTERING, SAMPLE FORMATION.

REFERENCES

[1] **Tant Zin Po, Tin Chzho, Pya Son Ko Ko, Paye Teyn Nainga,** Metodika sistemy raspoznavaniya obrazov s pomoshchyu samoorganizuyushchikhsya kart Kokhonena neyronnykh setey na osnove Matlab [The procedure of pattern recognition system using self-organizing maps of neural networks based on Mathlab], Internet-Journal 'Naukovedenie' No. 5(2013) [http://naukovedenie.ru/PDF/27tvn513.pdf].

[2] **D. Kumar, C.S. Rai, S. Kumar,** Face recognition using self-organizing map and principal component analysis, In: Proc. on Neural Networks and Brain, ICNNB 2005. 3 (2005) 1469–1473.

[3] **A.S. Panfilova,** Sistema testirovaniya intellekta na baze faktornykh modeley i samoorganizuyushchikhsya kart Kokhonena [The system of intellect testing based on factor models and self-organizing maps], Neyrokompyutery, razrabotka, primeneniye, No. 9 (2012) 6 –12.

[4] **K.A. Gushchin, S.A. Dolenko, S.A. Burikov, T.A. Dolenko,** Primeneniye algoritmov klasterizatsii i ponizheniya razmernosti dannykh v zadachakh analiza sostava mnogokomponentnykh rastvorov [An application of clustering algorithms and data dimensionality reduction to the problems on composition of multicomponent solutions], 13th Vserossiyskaya nauchnaya konferentsiya «Neyrokompyutery i ikh primeneniye». Tezisy dokladov [Abstracts], MGPPU, Moscow, 2015.

[5] **A.V. Novikov,** Neyrosetevyye metody resheniya zadach klasternogo analiza [Neural-network methods of solving the problems on cluster analysis] «Neyrokompyutery: razrabotka, primeneniye». No. 2 (2014) 48–53.

[6] **T. Kokhonen,** Samoorganizuyushchiyesya karty [Self-organizing maps], Moscow, Binom, Laboratoriya znaniy, 2008.

[7] **V.I. Gorbachenko,** Neyroinformatika. Konspekt lektsiy [Neural informatics, Lecture notes], Penza: Penzenskiy gosudarstvennyy pedagogicheskiy universitet, 2011.

[8] **I.V. Kovalev,** Intellektualnaya sistema prognozirovaniya zagryazneniya atmosfery [The intelligence system to forecast air pollution], Neyrokompyutery: razrabotka, primeneniye. No. 7 (2010) 62–66.

[9] **S. Khaykin,** Neyronnyye seti: polnyy kurs [Neural networks: a full course of study], 2-nd ed. Per. s angl. Moscow, ID "Wyliams", 2008.

[10] **D.E. Bestens, V.M. Van Den Berg, D. Vud,** Neyronnyye seti i finansovyye rynki: prinyatiye resheniy v torgovykh operatsiyakh [Neural networks and financial markets: decision of making in the trading], Moscow, TVP, 1997.

[11] **N.B. Paklin, V.I. Oreshkov,** Biznes-analitika: ot dannykh k znaniyam [Business intelligence: from the data to the knowledge], SPb., Piter, 2013.

[12] **R. Kalan,** Osnovnyye kontseptsii neyronnykh setey [Fundamental concepts of neural networks], Moscow, Wyliams, 2001.

[13] **Beale M.H., Hagan M.T., Demuth H.B.** Neural Network Toolbox User's Guide [http://www.mathworks.com/help/pdf_doc/nnet/nnet_ug.pdf] // MathWorks, Inc. 2014.

[14] **V.I. Gorbachev,** Seti i karty Kokhonena [Networks and self-organizing maps]. URL: http://gorbachenko.self-organization.ru/articles/Self-organizing_map.pdf (data obrashcheniya: 11.01.2016).

[15] **V.I. Dmitriyev,** Prikladnaya teoriya informatsii [Applied information theory]: Uchebnoye posobiye dlya studentov vuzov, Moscow, Vysshaya shkola, 1989.

[16] **A.A. Pastukhov,** Predicting the parameters of energy installations with laser ignition: neural network models, St. Petersburg State Polytechnical University Journal. Physics and Mathematics. No. 2 (218) (2015) 19–29.

THE AUTHORS

PASTUKHOV Aleksey A.
National Research University of Electronic Technology
5 Pass. 4806, Zelenograd, Moscow, 124498, Russian Federation
pastuhov1992@gmail.com

PROKOFIEV Alexander A.
National Research University of Electronic Technology
5 Pass. 4806, Zelenograd, Moscow, 124498, Russian Federation
aaprokof@yandex.ru