Московский Физико-Технический Институт
(Государственный Университет)
Факультет Управления и Прикладной Математики
Кафедра ‹‹Интеллектуальные Системы››

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
БАКАЛАВРА

‹‹Анализ пространственно-временных рядов в задаче классификации кортикограмм››

Выполнил:
студент 4 курса 274 группы
Задаянчук Андрей Игоревич
Научный руководитель:
к.ф-м.н., н.с. ВЦ РАН
Стрижов Вадим Викторович

Содержание

Аннотация: Решается задача построения модели для точной и устойчивой классификации физической активности человека по временным рядам. Исследуются модели из класса двухслойных нейронных сетей. В первой частичасти находится оптимальная полная матрица ковариации. Используется апрокимация Лапласа и модификация алгоритма Ньфтона-Раффсона для итеративного нахождения оценки оптимальной матрицы ковариации. Исследуются свойства полученной оценки матрицы ковариации. Данная матрица используется в методе Белсли отбора признаков. Во второй части рассматриваются модели с избыточно сложной структурой. Структура модели оптимизируется путем удаления из нее наборов параметров — нейронов. Для оптимизации структуры нейронной сети и обеспечения устойчивости предлагается алгоритм генетического типа. Новизна работы заключается в том, что вероятность удаления наборов параметров определяется дисперсией параметров. В вычислительном эксперименте модели, порождаемые предложенной стратегией, сравниваются по двум критериям качества: точности и устойчивости. Модели оптимизируются на выборках, полученных путем выделения признаков из временных рядов.

Ключевые слова: *классификация*, *нейронные сети*, *устойчивость*, *критерии прореживания*, *генетический алгоритм*, *метод Белсли*, *ковариационная матрица*, *многоклассовая классификация*.

1 Введение

В работе рассматривается два подхода к получению моделей, которые обладают

большей точностью и устойчивостью. Первая часть работы посвящена нахождению оптимальных гиперпараметров модели. Вторая часть работы посвящается исследованию методов построения нейронной сети оптимальной структуры.

В первой части рассматривается задача выбора модели многоклассовой классификации, которая наилучшим образом описывает данные. Для оценки качества приближения данных вводится функция ошибки, вид которой определяется статистическими предположениями о характере распределения зависимой переменной и вектора параметров модели многоклассовой классификации. Параметры задающие распределение вектора параметров называются гиперпараметрами.

Требуется выбрать оптимальную модель, которая максимизирует точность классификации на тестовой выборке и устойчивость модели, при ограничениях на сложность модели. [15]. При предположении о характере распределения параметров модели, нахождение оптимальной модели сводится к максимизации совместного правдоподобия модели и данных по всем гиперпараметрам и параметрам. Существует несколько методов оценивания совместного правдоподобия, таких как Монте Карло, аппроксимация Лапласа и другие[16].

В работе предлагается оценивать эти гиперпараметры путем максимизации совместного правдоподобия модели и аппроксимации Лапласа. Данный подход использовался для нахождения оптимальных гиперпараметров в случае, когда матрица ковариации диагональна [17]. Полученно выражение для связи гессиана функции ошибки в точке оптимальности параметров со значеним оптимальной матрицы ковариации. Для получения оптимальной матрицы ковариации предлагается воспользоваться итеративным процессом, похожим на классический итеративный метод оптимизации Ньютона-Рафсона.

Во второй части работы исследуются и сравниваются методы изменения размерности пространства параметров двухслойных нейронных сетей. При уменьшении размерности пространства параметров значительно уменьшается время оптимизации параметров, увеличивается обобщающая способность нейронной сети, и, как следствие, уменьшается значение функции ошибки на контрольной выборке [Помилка: джерело посилання не знайдено][1].

Оптимизировать размерность пространства можно на разных уровнях — на уровне

нейронов (наборов параметров) [2] и отдельных параметров [3]. Структурные параметры модели – это параметры, управляющие включением нейрона в модель. В данной работе размерность пространства параметров оптимизируется на уровне нейронов, путем изменения значений структурных параметров. Предложено несколько способов такой оптимизации: прореживание (network pruning) [4], наращивание (network growing) [5] и пошаговое чередование наращивания и прореживания [6, 7, Помилка: джерело посилання не знайдено]. В настоящей работе рассматривается оптимизация размерности с помощью прореживания. Базовыми алгоритмами прореживания нейронных сетей являются оптимальные прореживания (англ. «optimal brain damage»[8] и «optimal brain surgery»[9]), основанные на вычислении вторых производных функции ошибки.

Устойчивая и оптимальная модель описывается с помощью генетического алгоритма на уровне нейронов [10, 11] путем оптимизации структурных параметров. Базовые алгоритмы [8, 9] находят локальный минимум функции ошибки. В случае же, когда функция ошибки имеет значительное число локальных минимумов, найденный минимум может не совпадать с глобальным. Для нахождения глобального минимума используется алгоритм отбора моделей путем случайного подбора, комбинирования и вариации структурных параметров [10] набора нейронных сетей. Вероятность комбинирования и вариации структурных параметров нейронной сети тем меньше, чем больше показатель выпуклости, используемый в «optimal brain damage» [8].

В вычислительном эксперименте рассматривается задача классификации физической активности человека по измерениям акселерометра. Эта задача решалась в исследованиях [12, 13] с помощью нейронных сетей. В вычислительном эксперименте оцениваются значения критериев качества для нейронных сетей, порождаемых предложенной стратегией. Временные ряды предварительно обрабатываются двумя способами — экспертным порождением признаков [12] и вводом метрики выравнивания временных рядов, с последующим выделением признаков, полученных как расстояние до центроидов классов [14].

2 Нахождение оптимальной матрицы ковариации.

2.1 Постановка задачи

Дана выборка $D=\{(x_i,t_i)\}, i\in I=\{1\dots m\}$, состоящая из m объектов, каждый из которых описывается n признаками $x_i\in R^n$ и принадлежит одному из z классов $t_i\in\{0,1\}^z$. Задано разбиение множества индексов выборки $I=L\sqcup C$ на обучающую (x_l,t_l), где $l\in L$, и контрольную (x_c,t_c), где $c\in C$. Необходимо выбрать наиболее точную и при этом устойчивую модель классификации.

Моделью назовем отображение:

$$f:(w,X)\mapsto y, \tag{1}$$

$$w=[w_1,\dots,w_j,\dots,w_{mn}],\ \ j\in J=\{1,\dots,mn\},$$

где w — вектор параметров модели, $X\in R^{n\times k}$ — матрица объект-признак, $y\in[0,1]^m$ — зависимая переменная. Предполагается, что переменная y — мультиномиально распределенная случайная величина, а переменная w имеет нормальное распределение:

$$w:N(0,A), \tag{2}$$

A — ковариационная матрица.

В данной работе рассматриваются модели f, принадлежащие классу многоклассовой логистической регрессии с базисными функциями $\tanh$

$$a(x)=\Theta\tanh(x)=\Theta\varphi(x), \tag{3}$$

где Θ — это матрица весов $n\times m$, и

$$¿ m \exp\left(a_i\left(x\right)\right) ¿ ¿ , ¿$$

(4)

Вектор y интерпретируется как вектор вероятностей: y_i есть вероятность того, что вектор x принадлежит классу с номером i:

$$¿ m \, y_i = 1, \quad i = 1, \ldots, m . ¿$$

Под вектором параметров будем понимать $w = \mathbf{vec}\left(W\right).$ Модель (1) является суперпозицией функций (3),(4).

2.1.1 Максимизация правдоподобия модели

Рассмотрим совместное правдоподобие y, w:

$$p\left(y, w \middle| X, A\right) = p\left(y \middle| X, w\right) p\left(w \middle| A\right).$$

Представим правдоподобие модели, как интеграл от произведения правдоподобия данных и правдоподобия модели:

$$p\left(y \middle| X, w\right) p\left(w \middle| A\right) dw . ¿$$

(5)

Используя формулу полной вероятности и формулу (5) получаем задачу оптимизации матрицы A:

$$p\left(y \middle| X, w\right) p\left(w \middle| A\right) dw , ¿$$

где M^n — множество всех положительно определенных матриц размера $n \times n$.

2.2 Оценка ковариационной матрицы методом Лапласа

Рассмотрим интеграл (5) и обозначим интегрируемую функцию $Q(w|A)$:

$$e^{\ln Q(w|A)}dw¿ \tag{6}$$

Предположим, что $\hat{w}$ максимизирует логарифм правдоподобия модели,

$$\hat{w}=\underset{w\in R^n}{\operatorname{argmax}} \ln Q(w,\hat{A}), \tag{7}$$

где $\hat{A}$ — оценка ковариационной матрицы, максимизирующая (6). Тогда $\hat{A}$ — искомая оценка матрицы ковариации.

Аппроксимация Лапласа использует разложение $\ln Q(w)$ в окрестности оптимального значения вектора параметров $\hat{w}$:

$$\ln Q(w)\approx \ln Q(\hat{w})-\frac{1}{2}(w-\hat{w})H(w-\hat{w}),$$

где H — матрица Гессе (гессиан) функции $\ln Q(w)$,

$$H=\nabla\nabla \ln Q(w)|_{w=\hat{w}}.$$

Вместо максимизации интеграла (6) будем максимизировать приближенную функцию:

$$\exp\left(-\frac{1}{2}(w-\hat{w})H(w-\hat{w})\right)dw \to \max_{A\in M^n}.¿ \tag{8}$$

Теорема 1 *В предположении* $\hat{w}:N(0,A)$*, в задаче многоклассовой классификации матрица оптимизирующая правдоподобие данных в аппроксимации Лапласа имеет вид:*

$$A=\left(H\left(\hat{w}\right)^{-1}+\left(H\left(\hat{w}\right)^{-1}\right)^{T}-\hat{w}\,\hat{w}^{T}\right)^{-1},$$

где $H\left(\hat{w}\right)$ это значение гессиана функции ошибки в при оптимальных параметрах $\hat{w}$.

Доказательство. Заметим, что подыинтегральное выражение (8) представляет собой сомножитель функции плотности многомерного нормального распределения и интеграл можно вычислить, домножив на нормирующий множитель:

$$f\left(A\right)=Q\left(\hat{w}\right)\left(2\pi\right)^{\frac{n}{2}}\left|H^{-1}\right|^{\frac{1}{2}}\to\max_{A\in M^{n}}.$$

(9)

Логарифмируя (9), получаем оптимизационную задачу:

$$\ln f\left(A\right)=\ln Q\left(\hat{w}\right)+\frac{n}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left|H\right|\to\max_{A\in M^{n}},$$

(10)

где $\ln Q\left(\hat{w}\right)$ — логарифм правдоподобия в точке $\hat{w}$. Распишем его подробнее:

$$\ln Q\left(\hat{w}\right)=\ln p\left(y\left|X,\hat{w}\right.\right)-\frac{n}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left|A\right|-\frac{1}{2}\hat{w}\,A\,\hat{w}.$$

(11)

Перепишем (10), подставив (11):

$$\ln f\left(A\right)=\ln p\left(y\left|X,\hat{w}\right.\right)-\frac{1}{2}\ln\left|A\right|-\frac{1}{2}\hat{w}\,A\,\hat{w}-\frac{1}{2}\ln\left|H\right|\to\max_{A\in M^{n}}.$$

(12)

Далее воспользуемся леммой 1 и получим, что оптимальное значение матрицы A:

$$\hat{A}=\left(H\left(w\right)^{-1}+\left(H\left(w\right)^{-1}\right)^{T}-\mathbf{w}\mathbf{w}^{\boldsymbol{T}}\right)^{-1}.$$

(13)

Лемма 1 *Максимальное значение функции*

$$h\left(A\right)=\ln p\left(y\left|X,\hat{w}\right.\right)-\frac{1}{2}\ln\left|A\right|-\frac{1}{2}\hat{w}\,A\,\hat{w}-\frac{1}{2}\ln\left|H\right|$$

доставляется с помощью матрицы

$$\hat{A}=\left(H\left(w\right)^{-1}+\left(H\left(w\right)^{-1}\right)^{T}-\mathbf{w}\mathbf{w}^{\boldsymbol{T}}\right)^{-1}.$$

Доказательство. Для нахождения оптимальной матрицы A продифференцируем по A

$$\frac{\partial h(A)}{\partial A}=\frac{\partial}{\partial A}(\ln p(y|X,\hat{w})-\frac{1}{2}\ln|A|-\frac{1}{2}\hat{w}A\hat{w}-\frac{1}{2}\ln|H|). \tag{14}$$

Продифференцируем каждое слагаемое по A:

$$\frac{\partial \ln p(y|X,\hat{w})}{\partial A}=0,\quad \frac{\partial \ln|A|}{\partial A}=(A^{-1})^{T},\quad \frac{\partial \hat{w}A\hat{w}}{\partial A}=\hat{w}\hat{w}. \tag{15}$$

Для нахождения $\frac{\partial \ln|H|}{\partial A}$ вычислим $\frac{\partial H}{\partial a_{ij}}$:

$$\frac{\partial H}{\partial a_{ij}}=\frac{\partial(-\nabla\nabla \ln E(w)-A)}{\partial a_{ij}}=-\frac{\partial A}{\partial a_{ij}}$$

С помошью выражения для $\frac{\partial H}{\partial a_{ij}}$ найдем $\frac{\partial \ln|H|}{\partial a_{ij}}$:

$$(H^{-1}\frac{\partial H}{\partial a_{ij}})=-h_{ij^{-1}}-h_{ji^{-1}},¿ \tag{16}$$

$$\frac{\partial \ln|H|}{\partial A}=-H(w)^{-1}-(H(w)^{-1})\cdot \tag{17}$$

Для нахождения максимума приравняем:

$$\frac{\partial \ln f(A)}{\partial A}=0.$$

Используя (14)-(17), выразим A из (??):

$$\hat{A}=\left(H\left(w\right)^{-1}+\left(H\left(w\right)^{-1}\right)^{T}-\mathbf{ww}^{\boldsymbol{T}}\right)^{-1}$$

Для оптимизации по w при фиксированной матрице A воспользуемся итеративным методом Ньютона-Рафсона:

$$w^{\text{new}}=w^{\text{old}}-H^{-1}\nabla\ln Q\left(w^{\text{old}}\right) \tag{18}$$

Для задачи многоклассовой классификации:

$$¿z\,t_{ij}\ln\left(p_j\left(x_i,w\right)\right).¿ \tag{19}$$

Вычислим градиент логарифма правдоподобия :

$$\nabla\ln Q(w)=-\nabla E(w)-\nabla\frac{1}{2}w\,\mathbf{Aw}=-\nabla E(w)-wA,$$

где $\left(y_{ij}-t_{ij}\right)\varphi\left(x_i\right)¿$, $y_{ij}=y_j\left(\varphi\left(x_i\right)\right)$.

Теперь можно выписать гессиан функции:

$$H=\nabla\nabla\ln Q(w)=-\nabla\nabla\ln E(w)-A, \tag{20}$$

где $\nabla\nabla\ln E(w)$ — блочная матрица с $m\times m$ блоками:

$$y_{ij}\left(I_{jk}-y_{ik}\right)\varphi\left(x_i\right)\varphi^{T}\left(x_i\right)¿ \tag{21}$$

Заметим, что матрица $\nabla\nabla\ln E(w)$ в (20) зависит от w, поэтому на каждом шаге оптимизации (18) для нового вектора w ее нужно пересчитывать. Данный метод является модификацией метода наименьших квадратов с итеративным пересчётом весов (IRLS).

2.3 Статистическая оценка ковариационной матрицы

Пусть W — матрица реализаций оптимального вектора параметров $\hat{w}$, определенного в (7) и рассматриваемого согласно (2) как многомерная случайная величина. Пусть эта матрица имеет размерность $nm \times k$.

По определению ковариационная матрица оптимального вектора параметров $\hat{w}$ вычисляется как:

$$A = \operatorname{cov}(W) = E(WW) - E(W)E(W) = E(WW).$$

Последнее равенство выполняется в силу предположения о том, что математическое ожидание вектора параметров равно нулю: $E(\hat{w}) = 0$. По матрице реализаций W многомерной случайной величины $\hat{w}$ ковариационная матрица может быть оценена следующим образом:

$$A = \frac{1}{nm}\mathbf{W}\mathbf{W}^{\boldsymbol{T}},$$

где $n \cdot m$ — это число параметров в модели. Для нахождения матрицы W, необходимо k раз запустить алгоритм оптимизации 3, на различных подвыборках, получить $\hat{w}_i$ и объединить их в матрицу W.

2.4 Использование полученной матрицы ковариации в методе Белсли.

Полученную ковариационную матрицу можно использовать, например, для отбора признаков методом Белсли. Эта задача сводится к нахождению индекса признака, который больше всего коррелирует с другими признаками. Пусть дан набор активных параметров w_j, $j \in A$. Это параметры, которые не были удалены на предыдущих шагах алгоритма.

Используя сингулярное разложение матрицы B получим выражение для матрицы A:

$$A = (\mathbf{BB}) = (\mathbf{U\Lambda V^T V\Lambda} U^T) = (\mathbf{U\Lambda\Lambda} U) = \mathbf{U\Lambda^2} U^T.$$

Индексом обусловленности η_ζ назовём отношение максимального элемента $\lambda_{\max}$ матрицы Λ к ζ-ому по величине элементу λ_ζ этой матрицы:

$$\eta_\zeta = \frac{\lambda_{\max}}{\lambda_\zeta}.$$

Так как ковариационная матрица A неполноранговая, то некоторые значения индексов обусловленности неопределены.

Оценками дисперсии параметров будут диагональные элементы A:

$$\sigma(w_\zeta) = A_{\zeta\zeta}.$$

Долевой коэффицент $q_{\zeta j}$ определим как вклад j-го признака в дисперсию ζ

-го элемента вектора параметров w :

$$q_{\zeta j}=\frac{u_{\zeta j}^2 \lambda_{jj}^2}{\sigma(w_\zeta)}.$$

Находим индексы обусловленности и долевые коэффициенты для набора активных параметров A. Большие значения индексов обусловленности указывают на зависимость между признаками. Поэтому для нахождения параметра, отвечающего этому критерию прореживания, находим максимальный индекс обусловленности:

$$\hat{\zeta}=\underset{\zeta\in A}{\operatorname{argmax}}\,\eta_\zeta.$$

Затем находим максимальный долевой коэффициент, соответствующий найденному максимальному индексу обусловленности $\eta_{\hat{\zeta}}$:

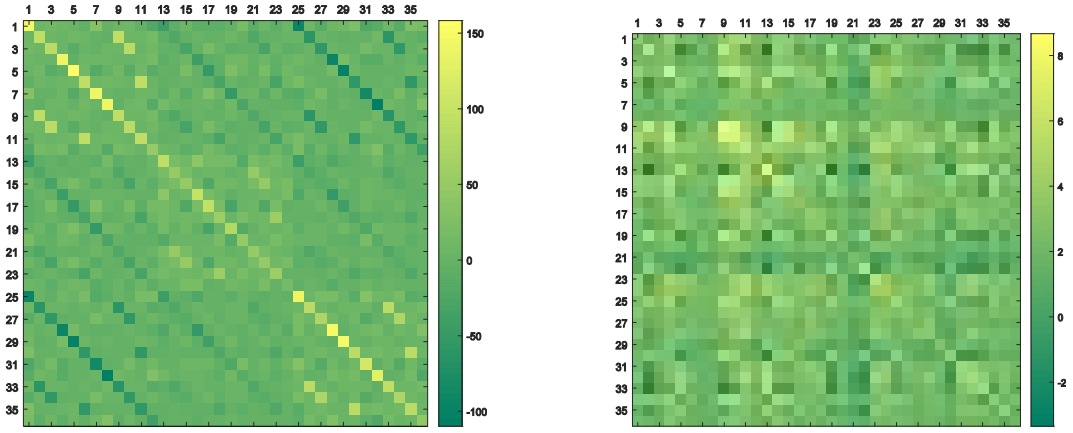
$$\hat{j}=\underset{j\in A}{\operatorname{argmax}}\,q_{\hat{\zeta}\,j}, \tag{22}$$

Параметр $w_{\hat{j}}$ и есть параметр, отвечающий критерию устойчивого прореживания.

2.5 Вычислительный эксперимент

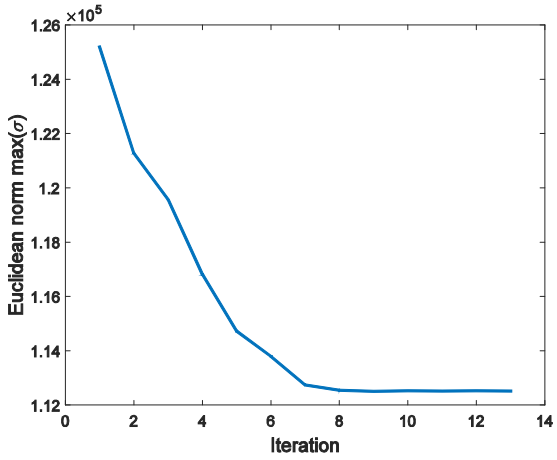
Вычислительный эксперимент проводился с двумя целями - сравнить полученную оценку матрицы ковариации полученную с помощью 3 с получаемой при статистической оценке матрицы ковариации и использользовать полученную оценку в алгоритме Белсли для отбора признаков.

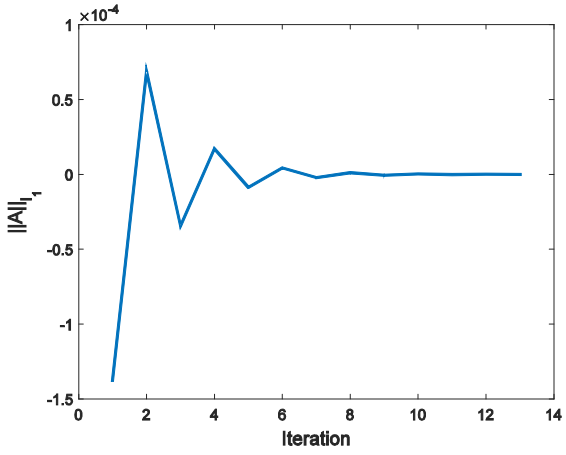
Был проделан вычислительный эксперимент на сгенерированной выборке, в которой часть признаков сильно коррелирует с другими. Размер сгенерированной выборки равен 10 000 объектов, каждый из которых состоит из 8 независимых признаков и 3 признаков, которые коррелируют с одним и 8 независимых . Объекты представляют собой 3 различных класса, каждый из которых сгенерирован из нормального распределения с центром в вершине 8-мерного гиперкуба и единичной дисперсией.



a) MIRLS b) Статистическая оценка

Рисунок 1: Оценка ковариационной матрицы





a) Норма $\|A\|_{\ell_1}$ b) Норма $\|A\|_{\ell_2}$

Рисунок 2: Сходимость оценки ковариационной матрицы

На Рис. 2 видно, что достаточно 10 итераций, чтобы ℓ_1 и ℓ_2 нормы матрицы стабилизировалась.

Из Рис. 3 видно, что функция ошибки (19) также сходится за несколько итераций.

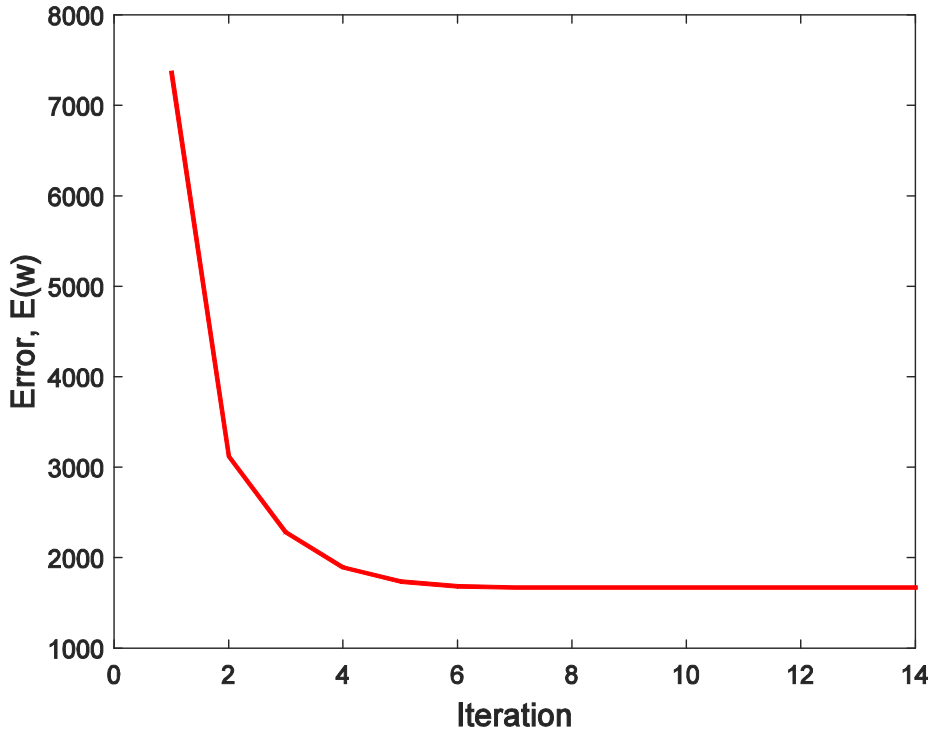


Рисунок 3: Сходимость функции ошибки $E(w)$

В качестве данных использовались данные из The Sleep-EDF Database [Expanded] [?]. Это открытая база данных которую можно найти на сайте Physionet.org [?]. Эта база содержит записи EEG 61 полисомнограмму, каждая из которых является набором записей EEG каналов людей, записанные во время их сна. Эти записи разбиты на 30-ти секундные отрезки, каждый из которых классифицирован экспертами. Классификация делалась на 5 классов (по системе AASM). Полученная выборка использовалась для классификации с помощью MIRLS. Результаты классификации отображенны в таблице 1.

Таблица 1. Показатели Presicion и Recall полученной модели

Класс	Presicion, %	Recall, %
W	30,6	39,4
1	82,8	91,3
2	84,5	69,9
3	60,8	66,4
R	84,3	69,4

3. Определение оптимальной структуры нейронной сети.

3. 1 Постановка задачи

Дана выборка $D=\{(x_i,t_i)\}, i\in I=\{1\dots m\}$, состоящая из m объектов, каждый из которых описывается n признаками $x_i\in R^n$ и принадлежит одному из z классов $t_i\in\{0,1\}^z$. Задано разбиение множества индексов выборки $I=L\sqcup C$ на обучающую (x_l,t_l), где $l\in L$, и контрольную (x_c,t_c), где $c\in C$. Необходимо выбрать наиболее точную и при этом устойчивую модель классификации.

Определение 1. *Моделью назовем отображение:*

$$f:(w,X)\mapsto y, \tag{1}$$

$$w=¿ w_1,\dots,w_j,\dots,w_J ¿^T, j\in J=\{1,\dots,J\},$$

где w — вектор параметров модели, $X\in R^{n\times m}$ — матрица объект-признак, $y\in\{0,1\}^z$ — зависимая переменная.

Предполагается, что переменная y — мультиномиально распределенная случайная величина, а переменная w имеет нормальное распределение:

$$w \;\; N(0,A^{-1}), \tag{2}$$

A^{-1} — ковариационная матрица. В данной работе рассматриваются модели f,

принадлежащие классу двуслойных нейронных сетей с функциями активации $\tanh$ и *softmax* .

$$a(x)=W_2^T\tanh(W_1^T x), \tag{3}$$

$$p(x)=\frac{\exp(a(x))}{\sum_l \exp(a_l(x))}. \tag{4}$$

Вектор p интерпретируется как вектор вероятностей: p_ξ есть вероятность того, что вектор x принадлежит классу с номером ξ :

$$p(x)=\{p_\xi\}, 0\le p_\xi\le 1, \sum p_\xi=1, \xi=1,\dots,z.$$

Под вектором параметров двуслойной нейронной сети будем понимать $w=(W_1^T\vee W_2^T)$, где W_1, W_2 — присоединённые матрицы весов первого и второго слоя нейронной сети. Вектор $y=\begin{matrix}y_1,\dots,y_\xi,\dots,y_z¿\\ ¿\end{matrix}^T$ определим следующим образом:

$$y_\xi=\begin{cases}1, & \text{если } \xi=\underset{\xi\in\{1,\dots,z\}}{argmax}(p_\xi),\\ 0, & \text{иначе}.\end{cases} \tag{5}$$

Моделью f является суперпозиция функций (3), (4), (5). В данной работе мы будем исследовать модели, отличающиеся друг от друга на структурном уровне. А именно содержащие разное количество групп связей — нейронов.

Определение 2. *Нейроном* u_k *назовем* k *-ю компоненту вектор-функции* $\tanh(W_1^T x)$ *— сомножитель (3).*

Каждый нейрон задается весами в соответствующей строке матрицы W_1^T .

Определение 3. *Нейрон назовем неактивным если* $u_k=0$.

Нейрон неактивен, если k строка матрицы W_1^T нулевая.

Определение 4. *Нейронной структурой* $A=\{k, u_k \neq 0\}$ *модели* f *назовем множество активных нейронов.*

Каждая нейронная структура A однозначно задает некоторую модель (1):

$$f_A : \hat{w}_A \in R^k,$$

где f_A — модель со структурой A, а $\hat{w}_A \in R^k$ — оптимальный вектор параметров модели f_A, определение которому будет дано ниже. Объединение всех f_A назовем множеством допустимых моделей:

$$F = ¿ A \subseteq J\{f_A\} \tag{6}$$

Оптимальную модель $\hat{f}_A$ будем выбирать из множества допустимых моделей f_A.

В качестве функции ошибки выберем функцию:

$$S(w \vee L) = -\sum_{i \in L} \sum_{\xi=1}^{z} t_{i\xi} \ln(p_\xi(x_i, w)). \tag{7}$$

Определение 5. *Устойчивостью* $\eta = \eta(\hat{w})$ *модели* f *с вектором параметров* w *назовем число* η *, равное числу обусловленности матрицы* A *, т.е.*

$$\eta(\hat{w}) = \frac{\lambda_{max}}{\lambda_{min}},$$

где λ_{max} *— максимальное, а* λ_{min} *— минимальное собственные числа матрицы* A.

Чем лучше обусловлена матрица A, тем более устойчива модель.

Матрица ковариации вычисляется с учетом предположения (2) о нулевом математическом ожидании вектора параметров w:

$$A^{-1} = cov(W) = E(W^T W) - E(W)E(W^T) = E(W^T W)$$

Где W - это матрица реализаций оптимального вектора параметров $\hat{w}$.

Определение 6. *Под точностью* S *модели* f *с вектором параметров* $\hat{w}$ *будем понимать значение функции ошибки (7) на контрольной выборке.*

Чем больше значение функции ошибки, тем меньше точность модели.

Определение 7. *Оптимальным вектором параметров модели* f_A *назовем такой вектор* $\hat{w}_A$ *, который является решением следующей задачи оптимизации:*

$$\hat{w}_A = \underset{w_A \in R^k}{argmin} S(w_A \vee L, f_A, \hat{A}). \tag{8}$$

Задача выбора оптимальной модели состоит в том, чтобы найти модель $f \in F$ для которой функция ошибки будет минимальной.

$$\hat{f}_A = \underset{f_A \in F}{argmin} S(f_A \vee C). \tag{9}$$

Устойчивость модели будет дополнительным критерием качества.

3.2 Описание алгоритма

Для получения оптимальной структуры модели в работе предлагается генетический алгоритм оптимизации структуры нейронной сети. Для сравнения также реализован базовый алгоритм NODE-OBD.

3.2.1 Алгоритм прореживания структуры нейронной сети NODE-OBD

Предлагаемый алгоритм определяет индекс нейрона, удаление которого приведет к минимизации приращения функции ошибки (7). Удаление нейрона эквивалентно занулению соответствующего столбца матрицы W_2, т.е. удалению сразу группы параметров вектора w. В этом разделе, для краткости изложения, будем обозначать матрицу W_2 как W. Предполагаем, что удаляемый нейрон наименьшим образом влияет на функцию ошибки. Для нахождения таких нейронов аппроксимируем функцию ошибки вблизи локального минимума матрицы W^0:

$$S(W_0+\Delta W)=S(W_0)+g^T(W_0)\Delta W+\frac{1}{2}\Delta W^T H\Delta W+O(\|\Delta W\|^3),$$

где ΔW — возмущение матрицы параметров в данной точке W_0; $g(W_0)$ — вектор градиента, вычисленный в точке W_0, $H=H(W_0)$ — матрица вторых производных функции ошибки. Предполагается, что функция ошибки S находится в окрестности локального минимума. Тогда ее аппроксимация записывается в следующем виде:

$$\Delta S=\frac{1}{2}\Delta W^T H\Delta W.$$

Пусть W_k — набор параметров соответствующий нейрону u_k, т. е. столбец

матрицы W , $W_k = W e_k$. Удаление этого нейрона (присвоение всем его параметрам нулевого значения) эквивалентно выполнению условия

$$\Delta W e_k + W_k = 0,$$

Получаем задачу условной минимизации

$$\Delta S = \frac{1}{2} \Delta W^T H \Delta W \to min, \Delta W e_k + W_k = 0.$$

Для решения этой задачи строим лагранжиан

$$L = \frac{1}{2} \Delta W^T H \Delta W - \lambda (\Delta W e_k + W_k).$$

Продифференцировав L по ΔW , получаем значение выпуклости L_k для элемента W_k :

$$L_k = \frac{W_k^T W_k}{2 [H^{-1}]_{k,k}}$$

где H^{-1} — матрица, обратная гессиану H ; $[H^{-1}]_{k,k}$ — k -ый диагональный элемент этой матрицы. Критерию оптимального прореживания отвечает группа параметров $W_{\hat{k}}$ с минимальным значением выпуклости:

$$\hat{k} = \underset{k \in A}{argmin} L_k.$$

Далее используя функцию выпуклости L_k как величину, определяющую вероятность комбинирования и вариации структурных параметров будет предложен недетерминированный вариант алгоритма «optimal brain damage».

3.2.2 Генетический алгоритм оптимизации структуры нейронной сети

Функция ошибки (7) является многоэкстремальной функцией вектора параметров w.

Поэтому при ее минимизации одним из детерминированных алгоритмом [Помилка: джерело посилання не знайдено, Помилка: джерело посилання не знайдено] определяется локальный минимум, который может не совпадать с глобальным минимумом. Для нахождения глобального минимума целесообразно воспользоваться недетерминированным генетическим алгоритмом.

Нейронная структура A задается бинарным вектором $a=[a_1,\ldots,a_K]$:

$$\begin{cases} a_q=1, & \textit{если } k\in A; \\ a_q=0, & \textit{иначе}. \end{cases}$$

Рассмотрим множество из M нейронных сетей с нейронными структурами $A_m, m=[1,\ldots,M]$, которому соответствует множество бинарных векторов $F_0=\{a_m\}, m=[1,\ldots,M].$ Назовем F_0 популяцией. Для каждого вектора a_m из множества F_0 оценивается вектор параметров $\hat{w}_{A_m}$ соответствующей нейронной сети f_{A_m} и вычисляется значение функции ошибки (7). Каждая из f_{A_m} оптимизирована алгоритмом обратного распространения ошибки [Помилка: джерело посилання не знайдено]. Опишем процедуру порождения новой популяции F_1 из популяции F_0 .

1. На множестве F_0 задается случайная величина θ , которая принимает значение a_m с вероятностью

$$p_m=\frac{\exp-\frac{Q_m}{Q_{max}}}{\sum_{l=1}^{N}\exp\frac{-Q_l}{Q_{max}}}\ , \qquad (10)$$

где суммарная выпуклость $Q_l=\sum_{k=1}^{K} L_k^{A_l}$ (здесь K – это число активных нейронов для

нейронной сети со структурой A_l) всех активных нейронов для нейронной сети со структурой A_l , а $Q_{max} = \max_{l \in \{1,\ldots,M\}} Q_l$.

Затем генерируется P реализаций случайной величины θ . Без ограничения общности будем считать, что P — четное число. Полученное множество векторов обозначим $F' = \{a_1^T, \ldots, a_P^T\}$.

2. Множество F' случайным образом разбивается на пары (a_s^T, a_t^T) , где $s, t = 1, \ldots, P, s \neq t$.

3. С каждой парой (a_s^T, a_t^T) производится операция скрещивания:

- генерируется случайное число $\zeta \in \{1, \ldots, K-1\}$;

- векторы (a_s^T, a_t^T) разделяются на две части и смешиваются следующим образом:

$$[a_s^1, \ldots, a_s^\zeta, a_t^{\zeta+1}, \ldots, a_t^K] \to a_{s'},$$

$$[a_t^1, \ldots, a_t^\zeta, a_s^{\zeta+1}, \ldots, a_s^K] \to a_{t'}.$$

4. С каждым вектором из F' проводится операция модификации:

- генерируется случайное число $\eta < K$;

- инвертируется значение позиций η вектора $a_{l'}$ и определяется вектор

$l\}$
$a_{¿}$.

Полученное множество векторов обозначается как $l\}\ \{\backslash\}\}\ \text{rsub}\ \{\text{l}=1\}\ \text{rsup}\ \{\text{P}\}$ $F_1=\{a_{¿}$ и является новой популяцией.

Таким образом, найден алгоритм для нахождения глобального минимума функции ошибки (7). Этот алгоритм использует выпуклость для определения вероятности использования структуры сети для комбинирования и вариации. Тем самым модели с меньшей выпуклостью будут с большей вероятностью использоваться для дальнейшего поиска оптимальной структуры сети.

3.3 Вычислительный эксперимент

Целью вычислительного эксперимента ставилось сравнение функции ошибки (7) генетического алгоритма оптимизации структуры нейронной сети с алгоритмом NODE-OBD.

Использовались данные с акселерометра мобильного телефона. Показания акселерометра записывались при шести видах физической активности: ходьба, бег, сидение, стояние, подъем и спуск. Далее эти показания обрабатывались экспертным порождением признаков и метрическим алгоритмом. При использовании временных рядов порождались следующие признаки: проекции среднего ускорения на координатные оси, среднеквадратические отклонения от проекций среднего ускорения на каждую из трех координатных осей, время между пиками синусоидального сигнала в миллисекундах.

В метрической выборке признаки были получены как расстояния до центроидов классов. Расстояние между рядами задавалось с помощью метода выравнивания. С методом выделения центроидов и введения расстояния между временными рядами можно ознакомиться в [Помилка: джерело посилання не знайдено].

В вычислительном эксперименте оптимизировалась двуслойная нейронная сеть с 40 нейронами в скрытом слое. Оптимизация проводилась по модифицированному OBD из раздела 3.1, а также по генетическому алгоритму из раздела 3.2. Для полученных на каждой итерации моделей (для лучшей модели в поколении популяции) подсчитана функция (7) по значению которой и сравнивалось качество моделей. На Рис. 4 сплошной линией

представлена зависимость функции ошибки (7) и пунктирной линией представлена зависимость устойчивости от числа итераций алгоритмов из разделов 3.1 и 3.2. (а) и (б) соответствуют функциям ошибки для генетического алгоритма и NODE-OBD для выборки полученной с помощью ручного выделения признаков. (в) и (г) соответствуют функциям ошибки для генетического алгоритма и NODE-OBD для метрической выборки.

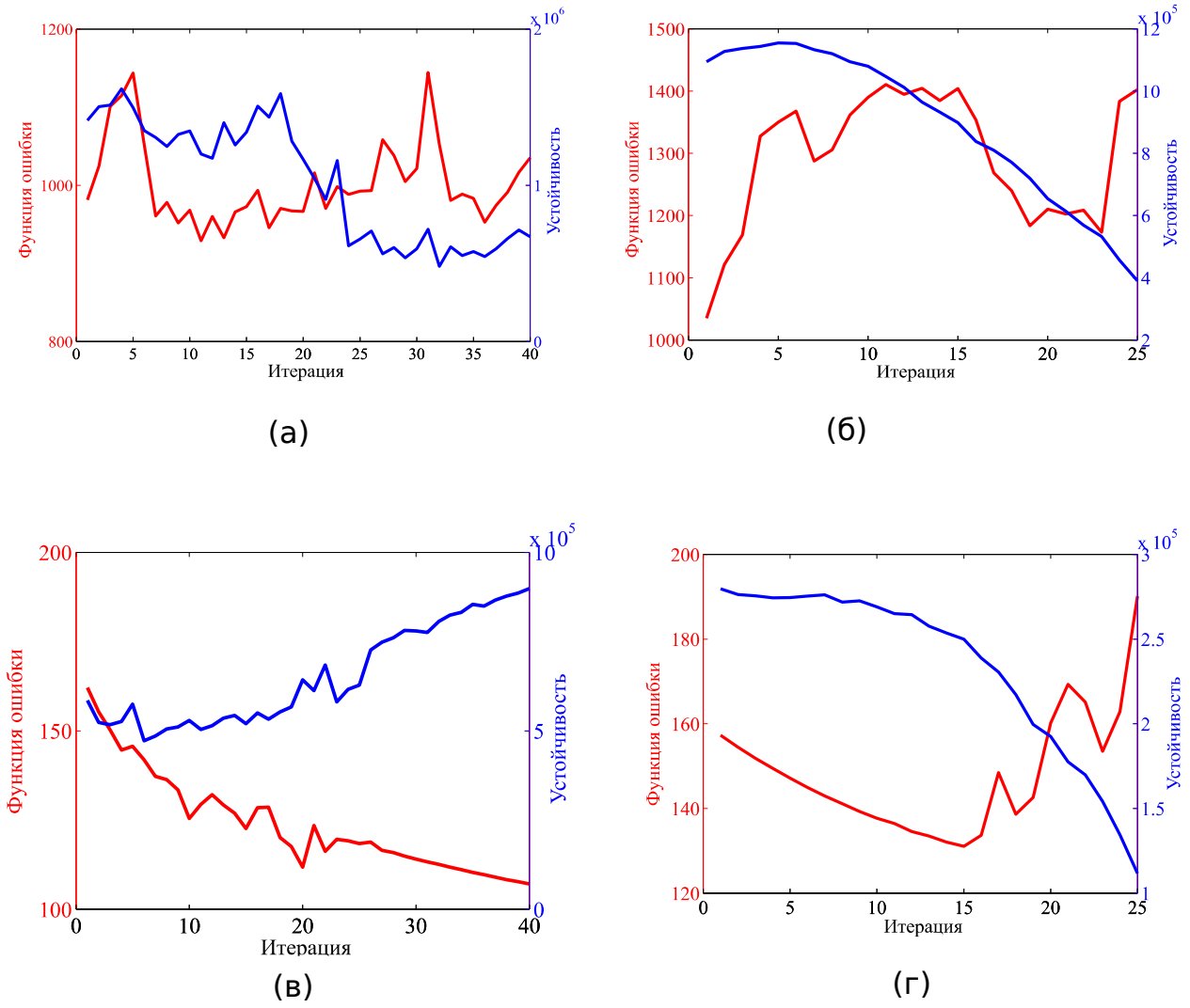
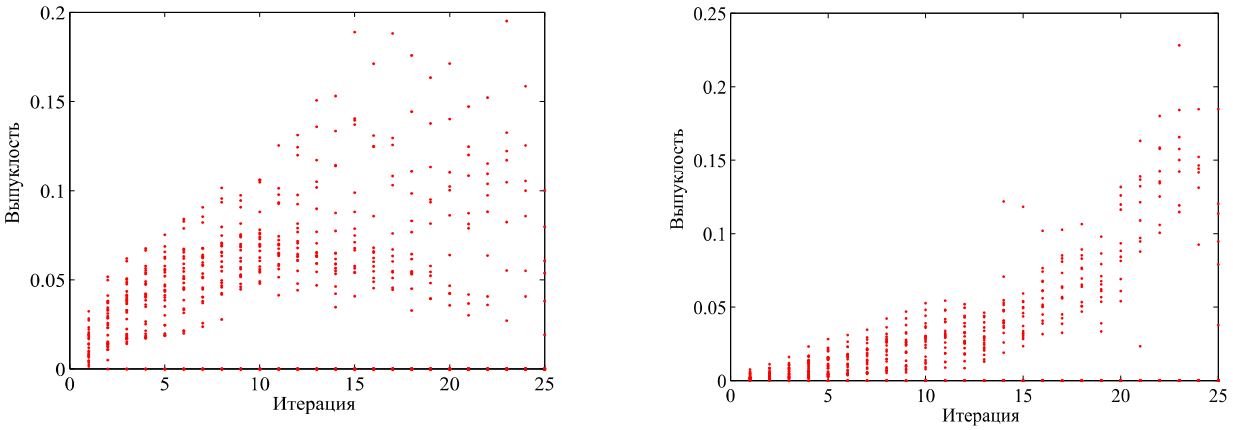


Рисунок 4. Зависимость значений функции ошибки от номера итерации.

Как видно из Рис. 4, генетический алгоритм позволяет получить большую точность за меньшее количество итераций для обоих выборок. При этом в случае экспертного выделения признаков устойчивость уменьшается, тогда как в метрической выборке она увеличивается. В случае базового алгоритма NODE-OBD функция ошибки (7) уменьшается незначительно или даже возрастает в случае экспертного порождения признаков. Устойчивость при этом в обоих случаях уменьшается.

На Рис. 5 отображено значение параметра выпуклости L_j (Salience) для всех активных нейронов. На каждой итерации количество активных нейронов уменьшается, а абсолютные значения выпуклости у оставшихся нейронов становятся больше по абсолютному значению, также увеличивается разность между значениями выпуклости



(б)

активных нейронов

Рисунок 5. Зависимость значений функции выпуклости нейронов сети от номера итерации:

(а) Экпертное порождение признаков

(б) Метрическая выборка

На Рис. 6 была визуализирована структура наиболее точной нейронной сети на каждой итерации. По горизонтали отложен номер итерации. Черная клетка означает, что нейрон активный, белая клетка — нейрон неактивный.

Представленные структуры, получены при оптимизации сети, на вход которой подавались выборка с выделенными вручную 43 признаками [Помилка: джерело посилання не знайдено] и выборка полученная с помощью метода выравнивания расстояния до центроидов классов.

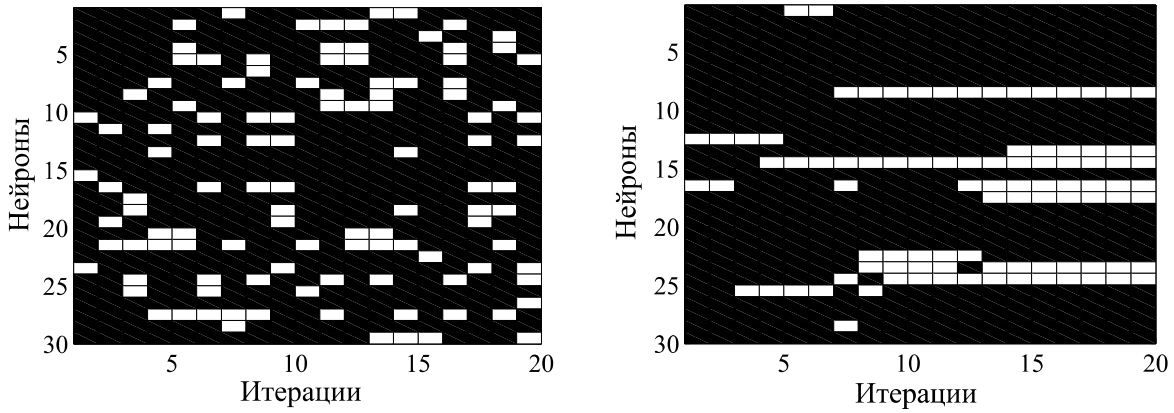


Рисунок 6. Изменение структуры оптимальной сети из популяции:

(а) Генетический алгоритм. Экспетрное порождение признаков

(б) Генетический алгоритм. Метрическая выборка

Для сравнения результатов, полученных с помощью исследуемых алгоритмов, с результатами других исследователей подсчитаны показатели Precision и Recall для наилучшей модели каждого из алгоритмов:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN},$$

где TP, FP, FN — это количество истинно-положительных, ложноположительных, ложноотрицательных объектов данного класса, соответственно. Эти показатели для каждого из классов занесены в таблицу 2. В ней отображены показатели для двух выборок для каждого из экспериментов.

Таблица 2. Показатели самых точных моделей каждого из алгоритмов

Вид физической активности	**OBD**				**Генетический алгоритм**			
	Первая выборка		**Метрическая выборка**		**Первая выборка**		**Метрическая выборка**	
	P,%	R,%	P,%	R,%	P,%	R,%	P,%	R,%
Бег	83,2	84,9	100	98,0	**88,3**	86,9	100	98,0
Ходьба	95,1	96,3	90,9	89,2	**98,0**	96,7	89,1	90,7
Подъем	51,1	46,5	85,1	83,3	55,3	51,0	85,1	90,9
Спуск	46,7	46,5	82,6	90,4	40,2	50,3	93,4	89,6
Сидение	92,5	91,3	98,1	98,1	90,4	89,4	98,1	98,1
Стояние	93,1	92,1	100	98,0	93,1	92,1	100	98,0

4 Заключение

В первой части работы получено выражение для полной матрицы ковариации. Предложен алгоритм для последовательного обновления значений оптимальных параметров и гиперпараметров. В вычислительном эксперименте исследована скорость сходимости данной оценки матрицы ковариации и функции ошибки. Также проведено сравнение данной оценки со статистической оценкой матрицы ковариации.

В второй части работы предложены два алгоритма оптимизации структуры нейронной сети – генетический алгоритм прореживания и алгоритм NODE-OBD. Эти алгоритмы сравнивались по значениям функции ошибки и устойчивости. Вычислительный эксперимент показал, что NODE-OBD позволяет значительно уменьшить количество активных нейронов, не увеличивая функцию ошибки модели, а генетический алгоритм позволяет получить модель с таким же количеством нейронов, как и NODE-OBD, при этом уменьшая значения функции ошибки модели. Проведено сравнение работы алгоритма для двух видов обработки временных рядов. Наиболее точные результаты, сравнимые с результатами [11], получаются при использовании метрической выборки [Помилка: джерело посилання не знайдено] и генетического алгоритма.

Литература

1. *Ghosh J, Tumer K*. Structural adaptation and generalization in supervised feed-forward networks // Jl. of Artificial Neural Networks, 1994. Vol.1. No.4. P. 431–458.

2. *Chung F.,Lee T.* A node pruning algorithm for backpropagation networks // Int. J. Neural Syst , 1992. Vol.3, No.3.P. 301–314.

3. Попова М.С., Стрижов В.В. Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2015. Т.9. Вып1. С.79–89.

4. *Suisse M.V., Thimm G., Fiesler E., Thimm G., Fiesler E.* Pruning of neural networks // Technical report, 2014.

5. *MacLeod C., Maxwell G.M.* Incremental evolution in ANNs: Neural nets which grow // Artif. Intell. Rev, 2001. Vol.16. No.3. P. 201–224.

6. *Vukovic N., Miljkovic Z.* A growing and pruning sequential learning algorithm of hyper basis function neural network for function approximation // Neural Networks, 2013. Vol. 46, P.210–226.

7. *Strijov V.V., Krymova E.V., Weber G.W.* Evidence optimization for consequently generated models // Mathematical and Computer Modelling, 2013.Vol.57. No.1-2. P.50–56.

8. *Cun Y.L., Denker J.S., Solla S.A.* Optimal brain damage // Advances in neural information processing systems, 1990. Vol.2. P. 598–605.

9. *Hassibi B., Stork D. G.* Second order derivatives for network pruning: Optimal brain surgeon // Advances in Neural Information Processing Systems, 1993. Vol.5. No.1. P. 164–172.

10. *Leung F.H., Lam H., Ling S., Tam P. K.* Tuning of the structure and parameters of a neural network using an improved genetic algorithm // IEEE Transactions on Neural Networks, 2003. Vol.14. No.1. P. 79–88.

11. *Oh I., Lee J., Moon B. R.*, Hybrid genetic algorithms for feature selection // IEEE Trans. Pattern Anal. Mach. Intell, 2004.Vol.26. No.11. P.1424–1437.

12. *Kwapisz J. R., Weiss G. M., Moore S.* Activity recognition using cell phone

accelerometers // SIGKDD Explorations, 2010. Vol.12. No.2. P. 74–82.
13. *Rasekh A., Chen C., Lu Y.* Human activity recognition using smartphone // Technical report, 2014.
14. Гончаров А. В., Стрижов В.В., Попова М.В. Метрическая классификация временных рядов с выравниванием относительно центроидов классов // Системы и средства информатики, 2015. Vol.1. №4.
15. Стрижов В.В., Задаянчук А. И., Попова М.С.. Выбор оптимальной модели классификации физической активности по измерениям акселерометра //*Информатика и её применения* , 9(1):79–89, 2015.
16. Токмакова А.А., Стрижов В.В. Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков. //Информатика и её применения, 6(4):66–75, 2012.