Выбор оптимальной модели классификации временных рядов

Андрей Игоревич Задаянчук

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н., В. В. Стрижов

15.06.2016 г.

Цель исследований

Определить оптимальную структуру нейронной сети на уровне групп связей с помощью недетерминированной структурной оптимизации. Сравнить полученную модель с моделью, полученной методом Белсли с использованием полной ковариационной матрицы.

Проблема

Существующие алгоритмы OBD (Optimal brain damage) находят только локальный минимум многоэкстремальной функции ошибки.

Решение

Предлагается использование недетерминированного алгоритма в предположении, что это позволит найти глобальный минимум и оптимальную структуру сети.

Литература

1. 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.
2. 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.
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.

Постановка задачи структурной оптимизации нейронной сети

Найдя оптимальную структуру сети, получим более точную и устойчивую модель.
Двухслойная нейронная сеть:

$$\mathbf{a}(\mathbf{x}) = \mathbf{W}_2\mathbf{tanh}(\mathbf{W}_1\mathbf{x}), \qquad \mathbf{p}(\mathbf{x}) = \frac{\mathbf{exp}(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))}.$$

Функция ошибки:

$$S(\mathbf{w}|\mathcal{L}) = \sum_{i\in\mathcal{L}}\sum_{\xi=1}^{z} t_{i\xi} \ln\bigl(p_\xi(\mathbf{x}_i, \mathbf{w})\bigr).$$

Оптимальная модель $\mathbf{f} \in \mathfrak{F}$ для которой функция ошибки будет минимальной:

$$\hat{\mathbf{f}}_{\mathcal{A}} = \operatorname*{argmin}_{\mathbf{f}_{\mathcal{A}}\in\mathfrak{F}} S(\mathbf{f}_{\mathcal{A}}|\mathcal{C}),$$

где $\mathcal{A}$ — множество индексов ненулевых строк матрицы $\mathbf{W}_1^{\mathsf{T}}$.

Постановка задачи оптимизации гиперпараметров

Найдем модель при котором совместное правдоподобие модели и данных максимально.
Двухслойная нейронная сеть:

$$\mathbf{a}(\mathbf{x}) = \mathbf{W}_2\mathbf{tanh}(\mathbf{W}_1\mathbf{x}), \qquad \mathbf{p}(\mathbf{x}) = \frac{\mathbf{exp}(\mathbf{a}(\mathbf{x}))}{\sum_j \exp(a_j(\mathbf{x}))}.$$

Пусть $\mathbf{w}$ — нормально распределенная случайная величина,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}).$$

Найдем ковариционную матрицу при котором правдоподобие данных максимально:

$$\mathbf{A}^* = \underset{\mathbf{A}\in\mathbb{M}^n}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{A}) = \underset{\mathbf{A}\in\mathbb{M}^n}{\operatorname{argmax}} \int_{\mathbb{R}^n} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w} \mid \mathbf{A})\, \mathrm{d}\mathbf{w}.$$

Оценка ковариационной матрицы методом Лапласа

Чтобы упростить выражение для правдоподобия, воспользуемся аппроксимацией Лапласа.

$$\int\limits_{\mathbb{R}^n} \underbrace{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w} \mid \mathbf{A})}_{Q(\mathbf{w}|\mathbf{A})} \, \mathrm{d}\mathbf{w} = \int\limits_{\mathbb{R}^n} e^{\ln Q(\mathbf{w}|\mathbf{A})} \, \mathrm{d}\mathbf{w}.$$

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

$$\ln Q(\mathbf{w}) \approx \ln Q(\hat{\mathbf{w}}) - \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^{\mathsf{T}} \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}).$$

Вместо максимизации интеграла будем максимизировать приближенную функцию:

$$Q(\hat{\mathbf{w}}) \int\limits_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^{\mathsf{T}} \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}}) \right) \mathrm{d}\mathbf{w} \to \max_{\mathbf{A} \in \mathbb{M}^n}.$$

Оптимальная матрица ковариации в приближении Лапласа

Теорема. Задаянчук, 2016

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

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

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

Доказательство теоремы

Домножим на нормирующий множитель и прологарифмируем:

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

Используя предположение о нормальности распределения $\hat{\mathbf{w}}$, получаем:

$$\ln Q(\hat{\mathbf{w}}) = \ln p(\mathbf{y} \mid \mathbf{X}, \hat{\mathbf{w}}) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}| - \frac{1}{2}\hat{\mathbf{w}}^{\mathsf{T}}\mathbf{A}\hat{\mathbf{w}},$$

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

Доказательство теоремы

$$\frac{\partial \ln f(\mathbf{A})}{\partial \mathbf{A}} = \frac{\partial}{\partial \mathbf{A}}(\ln p(\mathbf{y} \mid \mathbf{X}, \hat{\mathbf{w}}) - \frac{1}{2}\ln|\mathbf{A}| - \frac{1}{2}\hat{\mathbf{w}}^{\mathsf{T}}\mathbf{A}\hat{\mathbf{w}} - \frac{1}{2}\ln|\mathbf{H}|).$$

$$\frac{\partial \ln p(\mathbf{y} \mid \mathbf{X}, \hat{\mathbf{w}})}{\partial \mathbf{A}} = 0,$$

$$\frac{\partial \mathbf{H}}{\partial a_{ij}} = \frac{\partial(\nabla\nabla \ln p(\mathbf{y} \mid \mathbf{X}, \hat{\mathbf{w}}) - \mathbf{A})}{\partial a_{ij}} = -\frac{\partial \mathbf{A}}{\partial a_{ij}},$$

$$\frac{\partial \ln|\mathbf{H}|}{\partial a_{ij}} = \mathrm{tr}(\mathbf{H}^{-1}\frac{\partial \mathbf{H}}{\partial a_{ij}}) = -h_{ij}{}^{-1} - h_{ji}{}^{-1},$$

$$\frac{\partial \ln|\mathbf{H}|}{\partial \mathbf{A}} = -\mathbf{H}(\hat{\mathbf{w}})^{-1} - (\mathbf{H}(\hat{\mathbf{w}})^{-1})^{\mathsf{T}}.$$

Приравнивая $\frac{\partial \ln f(\mathbf{A})}{\partial \mathbf{A}} = 0$, выразим $\mathbf{A}$:

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

Модификация метода итеративно взвешиваемых наименьших квадратов (MIRLS)

Для нахождения оптимальных $\hat{\mathbf{w}}$ и $\hat{\mathbf{A}}$ воспользуемся итеративным методом:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \mathbf{H}(\mathbf{w}^{\text{old}})^{-1} \nabla \ln Q(\mathbf{w}^{\text{old}}),$$

$$\mathbf{A}^{\text{new}} = (\mathbf{H}(\mathbf{w}^{\text{new}})^{-1} + (\mathbf{H}(\mathbf{w}^{\text{new}})^{-1})^{\mathsf{T}} - \mathbf{w}^{\text{new}}\mathbf{w}^{\text{new}\mathsf{T}})^{-1},$$

где $\mathbf{H}(\mathbf{w}^{\text{new}}) = -\nabla\nabla \ln E(\mathbf{w}^{\text{new}}) - \mathbf{A}^{\text{old}}$.
Вычислим градиент логарифма правдоподобия:

$$\nabla \ln Q(\mathbf{w}) = -\nabla E(\mathbf{w}) - \nabla \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w} = -\nabla E(\mathbf{w}) - \mathbf{w}^{\mathsf{T}}\mathbf{A},$$

Выпишем гессиан $\mathbf{H}$:

$$\mathbf{H} = \nabla\nabla \ln Q(\mathbf{w}) = -\nabla\nabla \ln E(\mathbf{w}) - \mathbf{A}.$$

Метод Белсли

Использование полученной матрицы позволяет находить наиболее коррелированные признаки,

$$\mathbf{A} = (\mathbf{B}\mathbf{B}^\intercal) = (\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal) = (\mathbf{U}\mathbf{\Lambda}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal) = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\intercal,$$

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

Найдем максимальный индекс обусловленности:

$$\hat{\zeta} = \operatorname*{argmax}_{\zeta \in \mathcal{A}} \eta_\zeta.$$

Находим максимальный долевой коэффициент $\eta_{\hat{\zeta}}$:

$$\hat{j} = \operatorname*{argmax}_{j \in \mathcal{A}} q_{\hat{\zeta} j}.$$

Параметр $w_{\hat{j}}$ и есть параметр, который необходимо удалить.

Оптимальное структурное прореживание на уровне нейронов (OBD)

Необходимо найти индекс нейрона, удаление которого приведет к минимизации приращения функции ошибки:

$$\Delta S = \frac{1}{2}\Delta \mathbf{W}^{\mathsf{T}}\mathbf{H}\Delta \mathbf{W} \to \min, \qquad \mathbf{e}_j^{\mathsf{T}}\Delta \mathbf{W} + \mathbf{W}_j = 0.$$

Для решения этой задачи строим лагранжиан

$$L = \frac{1}{2}\Delta \mathbf{W}^{\mathsf{T}}\mathbf{H}\Delta \mathbf{W} - \lambda(\Delta \mathbf{W}\mathbf{e}_k + \mathbf{W}_k).$$

Продифференцировав L по $\Delta\mathbf{W}$, получаем значение L_k для $\mathbf{W}_k$:

$$L_k = \frac{\mathbf{W}_k^{\mathsf{T}}\mathbf{W}_k}{2[\mathbf{H}^{-1}]_{k,k}}.$$

Критерию оптимального прореживания отвечает набор параметров $\mathbf{W}_{\hat{k}}$ с минимальным значением выпуклости:

$$\hat{k} = \operatorname*{argmin}_{k\in\mathcal{A}} L_k.$$

Недетерминированная модификация оптимального структурного прореживания

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

$$\begin{cases} a_k = 1, & \text{если } k \in \mathcal{A}; \\ a_k = 0, & \text{иначе.} \end{cases}$$

Опишем процедуру порождения новой популяции $\mathcal{F}'' = \{\mathbf{a}''_m\}$ из популяции $\mathcal{F} = \{\mathbf{a}_m\}$.

1. Генерируется P реализаций $\mathbf{a}_m$ с вероятностью

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

2. Множество $\mathcal{F} = \{\mathbf{a}_1^\mathsf{T}, \ldots, \mathbf{a}_P^\mathsf{T}\}$ разбивается на пары $(\mathbf{a}_s^\mathsf{T}, \mathbf{a}_t^\mathsf{T})$,
3. Векторы $(\mathbf{a}_s^\mathsf{T}, \mathbf{a}_t^\mathsf{T})$ разделяются и смешиваются следующим образом:

$$[a_s^1, \ldots, a_s^\zeta, a_t^{\zeta+1}, \ldots, a_t^K] \to \mathbf{a}'_s,$$
$$[a_t^1, \ldots, a_t^\zeta, a_s^{\zeta+1}, \ldots, a_s^K] \to \mathbf{a}'_t.$$

4. Инвертируется значение позиций η вектора $\mathbf{a}'_l$ и получается вектор $\mathbf{a}''_l$.

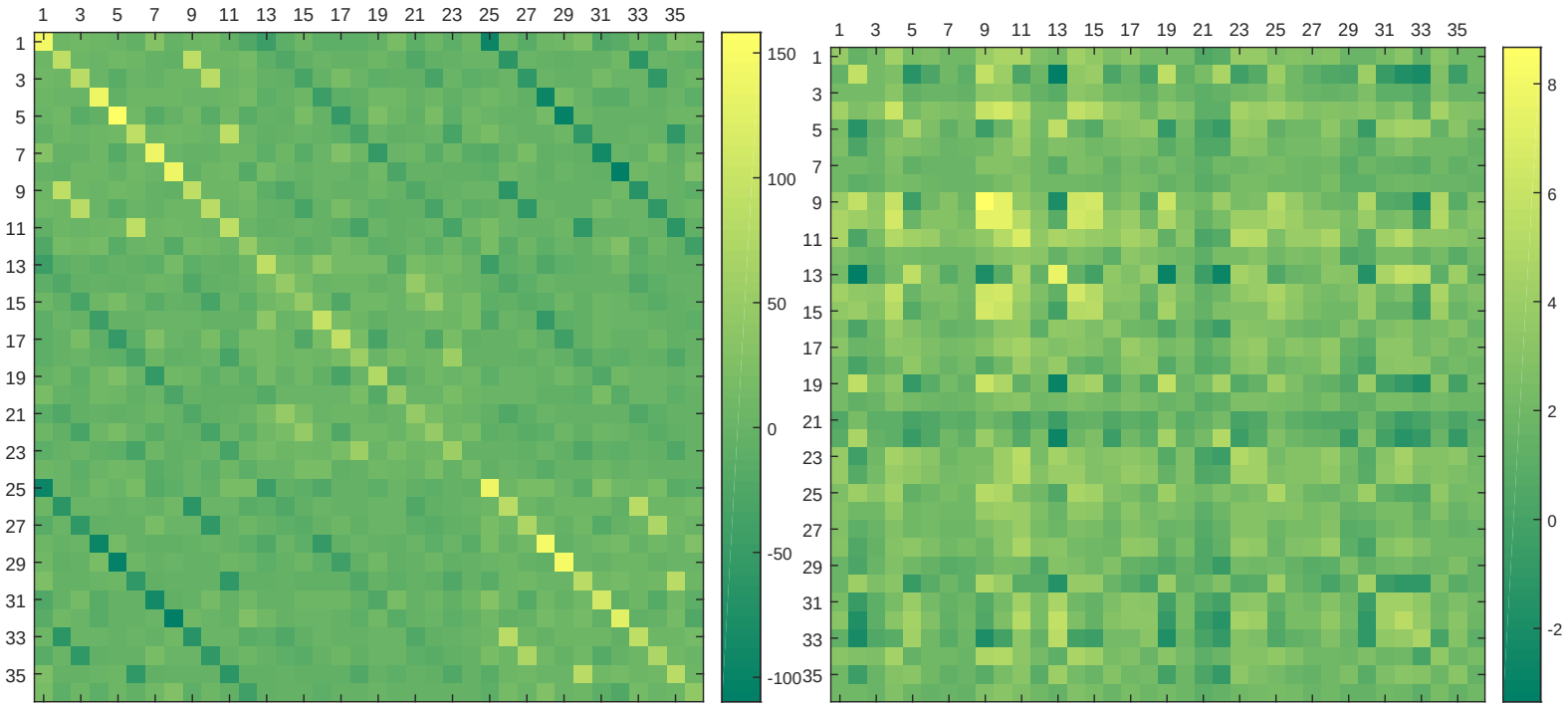
Полученное множество векторов и является новой популяцией.

Вычислительный эксперимент

- Использовались данные с акселерометра мобильного телефона. Всего было 4526 временных рядов.
- Шесть видов физической активности: ходьба, бег, сидение, стояние, подъем и спуск.
- 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. Всего было 42 признака.
- Kwapisz J. R., Weiss G. M., Moore S. Activity recognition using cell phone accelerometers // SIGKDD Explorations, 2010. Vol.12. No.2. P. 74–82.

Вычислительный эксперимент

Для матрицы полученной с помощью MIRLS характерно диагональное преобладание.

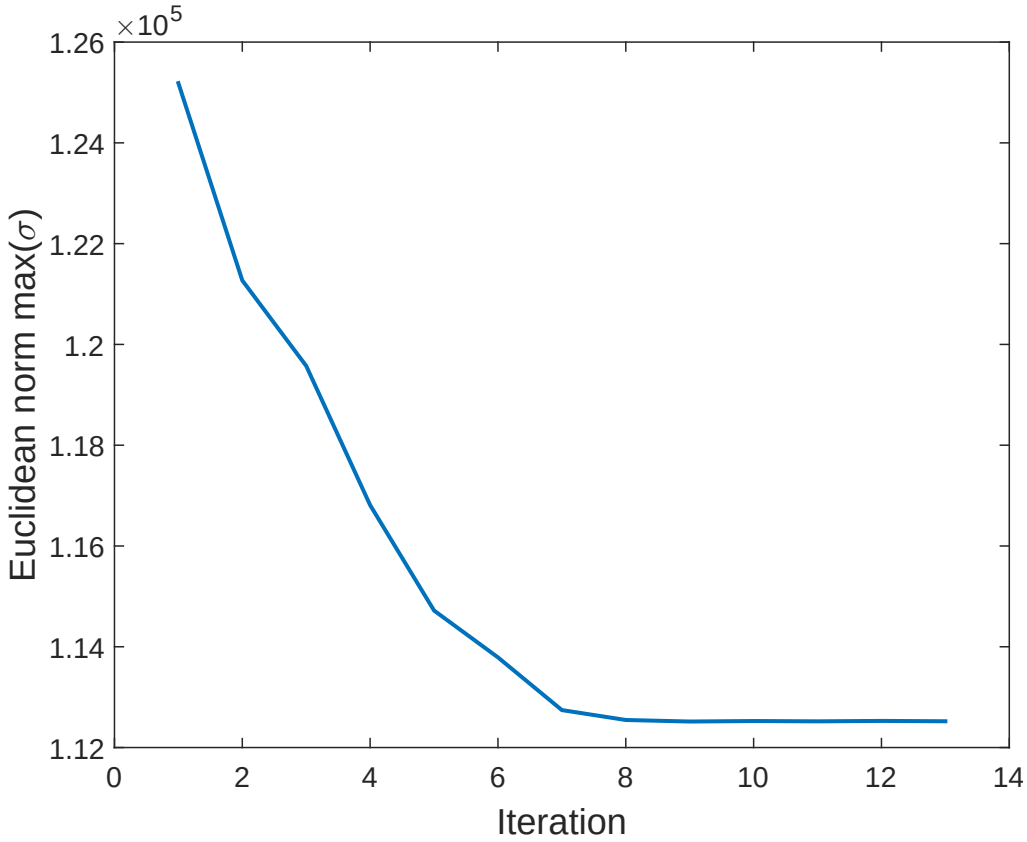


(а) MIRLS (б) Статистическая оценка

Различные способы оценки ковариационной матрицы

Вычислительный эксперимент

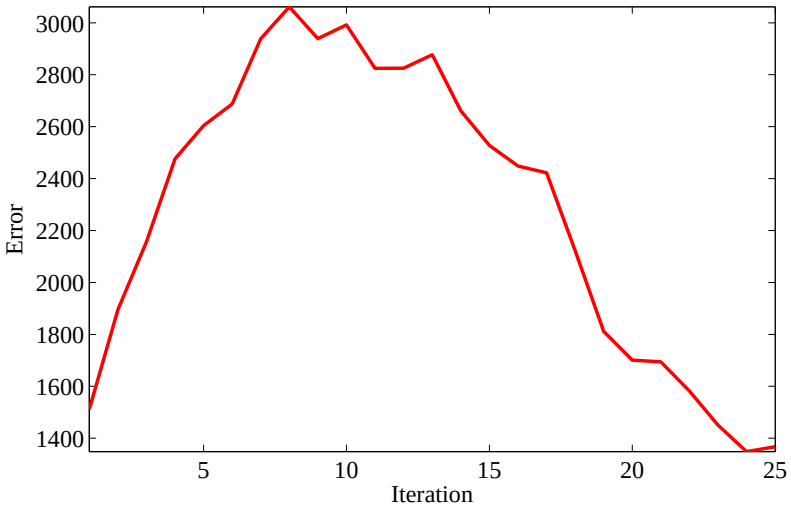
Достаточно 10 итераций для сходимости оценки ковариционной матрицы.



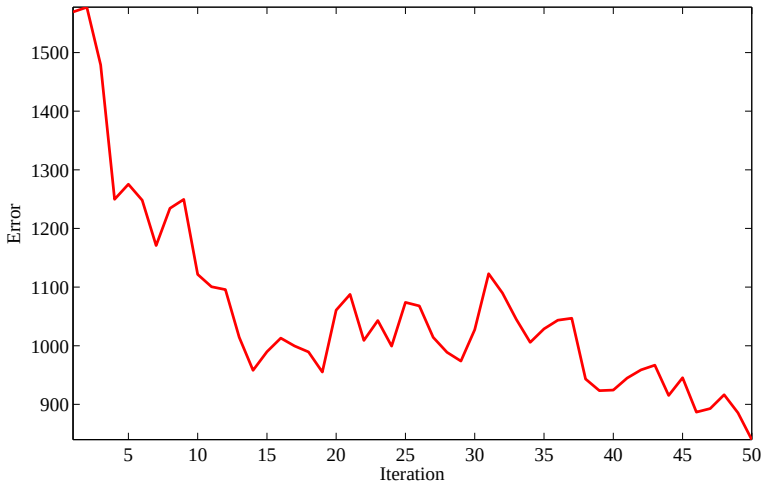
Сходимость оценки MIRLS ковариационной матрицы по эвклидовой норме

Вычислительный эксперимент

Генетический алгоритм позволяет значительно уменьшить функцию ошибки.



(в) OBD



(г) Генетический алгоритм

Зависимость функции ошибки от итерации

Вычислительный эксперимент

В среднем точность недетерминированной модификации OBD больше, чем у MIRLS.

Точность полученных моделей.

	Недетерминированный OBD		MIRLS	
Класс	Presicion, %	Recall, %	Presicion, %	Recall, %
Бег	88,3	86,9	75,4	95,1
Ходьба	98,0	96,7	93,7	96,1
Подъем	55,3	51,0	46,3	28,2
Спуск	40,2	50,3	30,4	8,5
Сидение	90,4	89,4	95,3	95,3
Стояние	93,1	92,1	94,3	94,3

Заключение

- Был разработан недетерминированный вариант OBD и показана его эффектность по сравнению с OBD
- Получен способ оценки оптимальной полной матрицы ковариации для задачи многоклассовой классификации
- Исследованны свойства этой матрицы
- Данная матрица использована для отбора признаков с помощью метода Белсли

Публикации:

- Задаянчук. А. И., Попова М. С., Стрижов В.В. Выбор оптимальной модели классификации физической активности по измерениям акселерометра // Информатика и ее применения, 2016
- Задаянчук. А. И., Стрижов В.В. Нахождение оптимальной матрицы ковариации для многоклассовой классификации временных рядов // Готовится к подаче