Московский государственный университет им. Ломоносова
Факультет Вычислительной математики и кибернетики
Кафедра Математических Методов Прогнозирования

Соболева Дарья Михайловна

Применение тематической модели классификации в информационном анализе электрокардиосигналов

Научный руководитель:
д. ф.-м. н. Воронцов Константин Вячеславович

Москва
2016

Содержание

1 Эксперимент №1

1.1 Описание эксперимента

Введем обозначения:

W^c – словарь терминов «метки классов».

$C = |W^c|$ – число различных классов документов.

W^{gram3} - словарь терминов «триграммы».

$W = W^c \cup W^{gram3}$ - общий словарь терминов.

D – коллекция текстовых документов (кардиограмм).

Тематическая модель классификации:

$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \varphi_{ct}\theta_{td}, \quad c \in W^c.$

Используемые метрики качества (на LOOCV):

1. Мера AUC - площадь под рок-кривой в координатах чувствительность-специфичность

$$AUC = \frac{1}{C} \sum_{c \in C} \frac{1}{|D_c||D'_c|} \sum_{d \in D_c} \sum_{d' \in D'_c} [p(c|d) > p(c|d')]$$

2. Мера LogLoss. Оценка уверенности классификатора

$$-\ln p(y_{true}|y_{pred}) = -(y_{true} \ln y_{pred} + (1 - y_{true}) \ln(1 - y_{pred}))$$

3. Перплексия по каждой отдельной модальности

$$L(\varphi, \theta) = \sum_{d \in D} \sum_{w \in W^{c,gram3}} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td}$$

$$P = exp(-\frac{1}{n}L)$$

n - длина коллекции в словах.

4. Разреженность матрицы φ по каждой отдельной модальности

$\varphi = p(w|t), \quad w \in W^c, W^{gram3}$

5. Разреженность матрицы $p(t|c)$

$$p(t|c) = \frac{p(c|t)p(t)}{p(c)}$$

$$p(t) = \sum_{d \in D} p(t|d)p(d) \qquad p(d) = \frac{1}{n_d} \qquad p(c) = \frac{1}{n_c}$$

Эксперименты проводятся на эталонной болезни «Хронический холецистит» (ХХЭ).

X — кардиограммы ($|X| = 372$)
X_m — кардиограммы больных($|X_m| = 224$)

Во множество исследуемых параметров классификатора входят:

- Число тем $|T|$
- Вес модальности «метки классов» τ

Рассматриваемые диапазоны изменения параметров:

$|T| \in range(C, 6C, 1)$
$\tau \in range(1, 1e5, 10)$.

1.2 Цель

Построение конкурентноспособной тематической модели классификации, подбор её параметров и стратегии регуляризации для достижения максимально возможной разреженности распределений $p(w|t), p(c|t), p(t|d)$.

1.3 Проведение эксперимента

1.3.1 Поиск оптимального числа тем ($|T|$)

1. $|T| = 2$

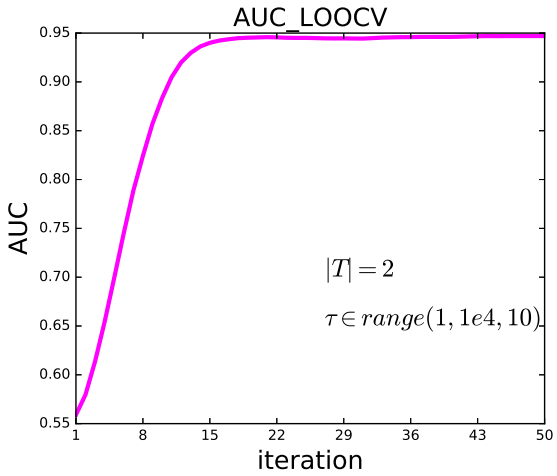

Рис. 1: AUC

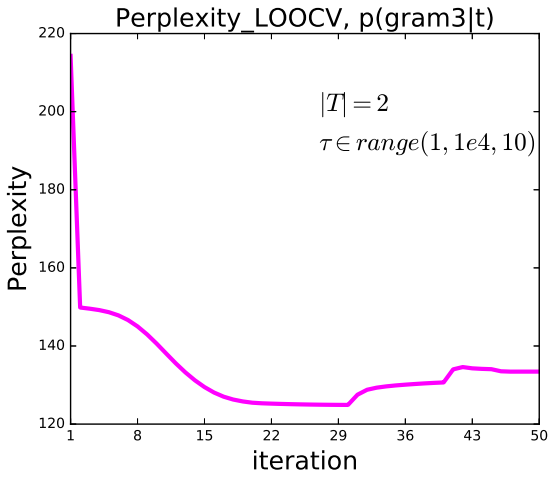

Рис. 2: Перплексия, триграммы

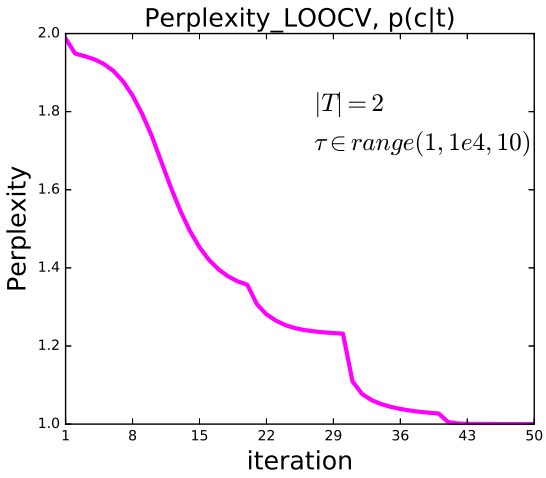

Рис. 3: Перплексия, метки классов

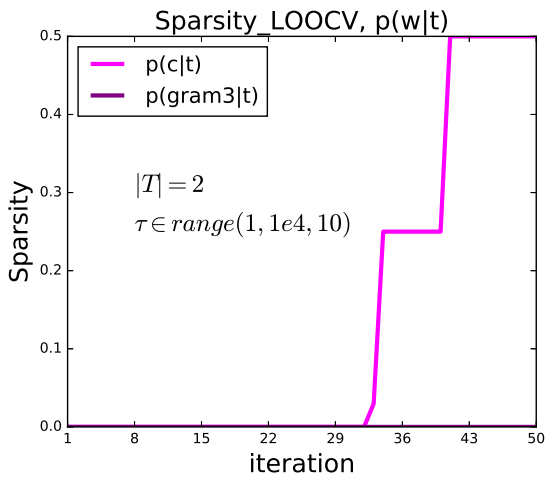

Рис. 4: Разреженность p(w|t)

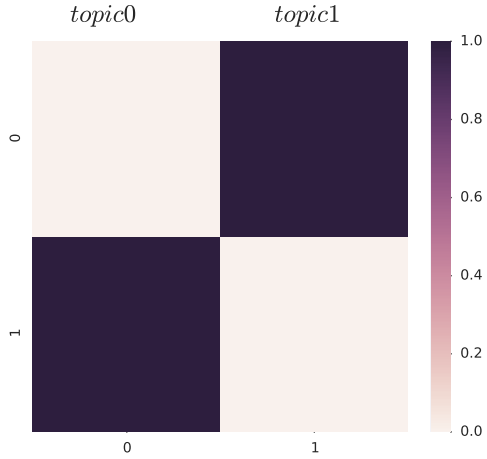

Рис. 5: $p(t|c)$