Московский государственный университет им. М. В. 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
2017

# Содержание

**Аннотация**

Технология информационного анализа электрокардиосигналов основана на преобразовании электрокардиограммы в последовательность символов. Ранее В. М. Успенским было показано, что с каждым заболеванием связан специфический набор трехсимвольных последовательностей (триграмм), называемых диагностическими эталонами. Также был предложен эвристический алгоритм поиска заданного числа диагностических эталонов для каждой конкретной болезни. В данной работе проверяется гипотеза о возможности автоматического выделения диагностических эталонов методами тематического моделирования.

# 1 Введение

Из физиологии сердца человека известно, что электрический, магнитный и гидродинамический импульсы, генерируемые сердцем во время его работы, являются источником важной информации о состоянии сердца и системы регуляции его функций. Электрофизиологические методы исследования и, в первую очередь, электрокардиография, получившая приоритетное развитие, в настоящее время играют важнейшую роль в современной кардиологии, в научной и практической медицине. Они позволяют достаточно глубоко оценить состояние миокарда и функций сердца.

Опыт изучения вариабельности сердечного ритма (ВСР) на основе длительной регистрации электрокардиограммы свидетельствует о том, что электрокардиоимпульсы могут быть носителями информации также о состоянии системы регуляции основных функций организма в норме, при различных заболеваниях и в условиях воздействия на человека экстремальных факторов профессиональной деятельности и среды обитания [1, 2]. Сердце генерирует импульсы электрической, магнитной и гидродинамической природы под влиянием сложного комплекса компонентов системы регуляции. Эти импульсы распространяются в объёме всего организма и имеют свойства сигналов. Вариации сердечного ритма носят иррегулярный характер, однако изолированное сердце вне организма человека генерирует импульсы без какой-либо вариабельности.

На основе этих наблюдений в [3] предлагается теория информационной функции сердца и обосновывается роль сердца как информационного органа. Предположение о том, что в организме существуют механизмы передачи сигналов, аналогичные амплитудной и частотной модуляции в теории сигналов и связи, приводит

к идее исследования вариабельности не только R-R-интервалов, но и R-амплитуд. Поиск способов демодуляции этих сигналов и дешифровки содержащейся в них информации о заболеваниях привёл к созданию технологии информационного анализа электрокардиосигналов [3, 4, 5, 6, 7]. Исследования в этом направлении проводились на базе Военно-медицинской академии (г. Санкт-Петербург), Российской медицинской академии последипломного образования, Российской академии космонавтики имени К. Э. Циолковского (ЦНИБИП), Государственного института усовершенствования врачей МО РФ и 2-го Центрального военного клинического госпиталя имени П. В. Мандрыка МО РФ, ныне Медицинского учебно-научного клинического центра имени П. В. Мандрыка.

В настоящее время технология информационного анализа электрокардиосигналов реализована в диагностической системе «Скринфакс» [3]. Она позволяет диагностировать по одной электрокардиограмме более 30 различных заболеваний внутренних органов, не ограничиваясь заболеваниями сердечно-сосудистой системы. Была накоплена обучающая выборка более 20 тысяч прецедентов – записей электрокардиограмм и соответствующих им диагнозов.

Технология информационного анализа электрокардиосигналов основана на преобразовании каждой электрокардиограммы сначала в последовательность интервалов и амплитуд кардиоциклов, затем в символьную последовательность – кодограмму и, наконец, в числовой вектор – частотное описание встречаемости трехсимвольных последовательностей (триграмм). В. М. Успенский показал, что существуют наборы триграмм, совместно встречающихся у определенной группы больных, но никогда совместно не встречающихся у здоровых людей. Эти наборы были названы диагностическими эталонами, поиск которых производился экспертом с помощью специально разработанной программы статистического анализа выборки кодограмм.

В данной работе приводится автоматизированный алгоритм выделения диагностических эталонов при помощи методов тематического моделирования. Осуществляется построение диагностического правила тематической модели классификации. Приводится сравнительный анализ конкурирующих методов на каждом из 20 различных заболеваний внутренних органов.

# 2 Технология информационного анализа ЭКГ-сигналов

## 2.1 Вариабильность сердечного ритма

Анализ вариабельности сердечного ритма (ВСР) основан на определении последовательности R-R-интервалов электрокардиограммы. Одним из подходов кодирования найденной последовательности является метод символьной динамики, впервые установленный в 1905 году [8]. Статистические показатели рассчитывались из анализа символьных последовательностей, полученных с помощью кодирования ряда R-R-интервалов. Было предложено два способа 4-буквенного кодирования: по величине отклонения текущего значения от среднего значения интервалов, а также на основе разности между соседними величинами интервалов. В технологии информационного анализа электрокардиосигналов каждый R-R-интервал кодируется тройкой $(R_n, T_n, \alpha_n)$, где $R_n$ – амплитуда, $T_n$ – величина интервала, $\alpha_n$ – аналог фазового угла в гармонических сигналах, определяемый как арктангенс отношения амплитуды к величине интервала: $\alpha_n = \operatorname{arctg} \frac{R_n}{T_n}$. Другие способы кодирования и вычисляемые характеристики исследуются в работах [9, 10, 11].

## 2.2 Дискретизация

В. М. Успенским было установлено, что для диагностики заболеваний важны знаки приращений амплитуд $R_n$, интервалов $T_n$ и углов $\alpha_n$. Возможны только 6 сочетаний увеличений и уменьшений этих трех величин, которые кодируются первыми буквами 6-символьного алфавита $\mathcal{A} = \{A, B, C, D, E, F\}$:

$$
\begin{aligned}
s_n = A: &\qquad R_n < R_{n+1}, &\qquad T_n < T_{n+1}, &\qquad \alpha_n < \alpha_{n+1} \\
s_n = B: &\qquad R_n \geqslant R_{n+1}, &\qquad T_n \geqslant T_{n+1}, &\qquad \alpha_n < \alpha_{n+1} \\
s_n = C: &\qquad R_n < R_{n+1}, &\qquad T_n \geqslant T_{n+1}, &\qquad \alpha_n < \alpha_{n+1} \\
s_n = D: &\qquad R_n \geqslant R_{n+1}, &\qquad T_n < T_{n+1}, &\qquad \alpha_n \geqslant \alpha_{n+1} \\
s_n = E: &\qquad R_n < R_{n+1}, &\qquad T_n < T_{n+1}, &\qquad \alpha_n \geqslant \alpha_{n+1} \\
s_n = F: &\qquad R_n \geqslant R_{n+1}, &\qquad T_n \geqslant T_{n+1}, &\qquad \alpha_n \geqslant \alpha_{n+1}
\end{aligned}
$$

В результате дискретизации амплитудограмма и интервалограмма преобразуются в символьную последовательность $S = (s_n)_{n=1}^{N-1}$ состоящую из символов алфавита $\mathcal{A}$ и называемую кодограммой. Каждые 3 последовательных символа $(s_n, s_{n+1}, s_{n+2})$

образуют триграмму. В [3] показано, что существуют наборы триграмм, совместная встречаемость которых говорит о наличии у обследуемого определённого заболевания. Обнаружение такого набора может произойти на любой стадии заболевания, в том числе задолго до возникновения симптомов и перехода заболевания в активную фазу. Его наличие говорит о предрасположенности к заболеванию, и потому может применяться с целью ранней диагностики.

## 2.3 Векторизация

Частота триграммы $w$ определяется как отношение ее числа вхождений $n_w(S)$ в кодограмму $S$ к общему числу триграмм в кодограмме, равному $N-3$:

$$n_w(S) = \sum_{n=1}^{N-3} \prod_{j=0}^{2} [s_{n+j} = w_j], \quad p_w(S) = \frac{n_w(S)}{N-3}.$$

В технологии информационного анализа электрокардиосигналов используются 216-мерные векторы частот триграмм. Поэтому каждую триграмму можно интерпретировать как слово из словаря $W$, содержащего $6^3 = 216$ слов. Набор триграмм, описывающий отдельную кодограмму – как документ, а выборку кодограмм – как коллекцию текстовых документов.

## 2.4 Обучающая выборка

Задача построения диагностического правила по выборке больных и здоровых людей ставится как задача классификации с пересекающимися классами, поскольку у одного человека может быть несколько заболеваний. Обозначим через $y_0$ класс здоровых людей, через $y_1, ..., y_K$ – классы больных $K$ различными заболеваниями. Для каждой кодограммы из обучащей выборки $X = \{S_1, \ldots, S_l\}$ известно множество классов $Y(S_i)$. Если человек здоров, то для его кодограммы $Y(S_i) = \{y_0\}$, если же у него имеются заболевания, то $y_0$ не входит в $Y(S_i)$. По каждому заболеванию $y_k$ была отобрана выборка $X_k$ обследуемых с диагнозами, подтвержденными лабораторными и инструментальными методами даигностики. В свою очередь выборка $X_0$ представляет собой здоровых людей, не имеющих существенных отклонений от состояния нормы («абсолютно здоровые»). В экспериментах для каждой болезни строилась двухклассовая выборка, состоящая из кодограмм больных $X_k$ конкретным заболеванием и, множества кодограмм $X_0$ «абсолютно здоровых» людей. Перечень болезней,

по которым были собраны обучающие выборки представлен в таблице 1.

| Болезнь | | МКБ-10 | $\lvert X_k \rvert$ |
|---|---|---|---|
| Абсолютно здоровые | АЗ | | 261 |
| Миома матки | ММ | D25 | 761 |
| Холецистит хронический | ХХ | K81.1 | 523 |
| Вегетососудистая дистония | ВСД | F45.3 | 688 |
| Гастродуоденит | ГД | K29 | 428 |
| Язвенная болезнь | ЯБ | K25-28 | 766 |
| Гипертоническая болезнь | ГБ | I11 | 1950 |
| Ишемическая болезнь средца | ИБС | I20 | 1303 |
| Сахарный диабет | СД | E10-E11 | 822 |
| Мочекаменная болезнь | МКБ | N20 | 840 |
| Желчнокаменная болезнь | ЖКБ | K80 | 383 |
| Узловой зоб щитовидной железы | УЩ | D34 | 730 |
| Хронический гастрит | ХГ | K29 | 731 |
| Дискинезия желчевыводящих путей | ДЖВП | K83 | 714 |
| Рак общий | РО | C00-C97 | 448 |
| Аденома простаты | ДГПЖ | N40 | 335 |
| Аднексит хронический | АХ | N70 | 334 |
| Анемия | ЖДА | D50 | 562 |
| Асептический некроз головки бедренной кости | НГБК | M91.1 | 531 |
| Туберкулез | ТБЦ | A15-19 | 719 |
| Остеопороз | ОП | M80-82 | 244 |

Таблица 1: Для каждого заболевания: название, аббревиатура, код МКБ-10, объём выборки $X_k$.

# 3 Модели классификации

## 3.1 Линейные байесовские классификаторы

Линейными моделями классификации называются диагностические правила, в которых решение принимается по значению взвешенной суммы признаков:

$$a_k(S) = [b_k(S) \geqslant \beta_k], \qquad b_k(S) = \sum_{w \in W} \gamma_{kw} f_w(S),$$

где $b_k(S)$ – дискриминантная функция, оценивающая степень принадлежности кодограммы $S$ классу $y_k$, $\beta_k$ – порог принятия решения и $f_w(S)$ – числовой признак, монотонно зависящий от частоты триграммы $w$ в кодограмме $S$. Задача обучения диагностического правила состоит в том, чтобы по выборкам прецедентов двух классов,

здоровых $X_0$ и больных $X_k$, оптимизировать веса признаков $\gamma_{kw}$.

Согласно байесовской теории классификации, минимальным риском потерь обладает оптимальный байесовский классификатор вида:

$$a_k(S) = \left[ \ln \frac{\pi_k(S)}{\pi_0(S)} \geqslant \beta_k \right], \tag{1}$$

где $\pi_k(S)$ – модель плотности распределения класса $y_k$, порог $\beta_k$ зависит от соотношения количества ошибок на объектах класса больных $y_k$ и здоровых $y_0$.

Предположим, что триграммы, характерные для данного заболевания, появляются в кодограмме независимо друг от друга. Тогда частоты триграмм $p_w(S)$ являются независимыми случайными величинами. В этом случае многомерная плотность распределения $\pi_k(S)$ представляется в виде произведения одномерных плотностей, классификатор приобретает особо простой вид и называется *наивным байесовским классификатором*. Чтобы найти одномерные плотности, предположим, что появления одной и той же триграммы в кодограмме независимы друг от друга. Тогда число появлений $n_w(S)$ триграммы $w$ в кодограмме $S$ описывается распределением Пуассона, а плотность $\pi_k(S)$ представляется в виде произведения распределений Пуассона:

$$\pi_k(S) = \prod_{w \in W} \frac{\lambda_{kw}^{n_w(S)}}{n_w(S)!} \exp(-\lambda_{kw}),$$

где $\lambda_{kw}$ – параметр распределения Пуассона. Его несмещённая выборочная оценка $\lambda_{kw} = (N-3)F_w(X_k)$, где $F_w(X_k) = \frac{1}{|X_k|} \sum\limits_{S \in X_k} p_w(S)$ совпадает со средним числом вхождений триграммы $w$ в кодограммы класса $y_m$ [12]. Подставляя эти оценки в плотности $\pi_k(S)$ и далее в формулу (1), получим, что оптимальный байесовский классификатор является линейным с коэффициентами $\gamma_{kw}$, которые легко вычисляются по обучающей выборке:

$$b_k(S) = \sum_{w \in W} \gamma_{kw} p_w(S), \quad \gamma_{kw} = \ln \frac{F_w(X_k)}{F_w(X_0)}.$$

Альтернативный вариант наивного байесовского классификатора [12] основан на предположении, что диагностическую ценность имеют не частоты триграмм в кодограмме, а только то, какие триграммы встречаются чаще, чем они могли бы встречаться чисто случайно. Будем полагать, что встречаемости триграмм в каждом классе $y_k$ – независимые биномиальные случайные величины $f_w(S) = [p_w(S) \geqslant \theta]$ с пара-

метрами вероятности события $\mu_{kw} = P\{p_w(S) \geqslant \theta | S \in X_k\}$. Тогда

$$\pi_k(S) = \prod_{w \in W} \mu_{kw}^{[p_w(S) \geqslant \theta]} (1 - \mu_{kw})^{[p_w(S) < \theta]}.$$

Несмещённая выборочная оценка параметра $\mu_{kw}$ совпадает со значением встречаемости триграммы $w$ в выборке прецедентов класса $y_k$: $\mu_{kw} = B_w(X_k, \theta)$, где $B_w(X_k, \theta) = \frac{1}{|X_k|} \sum\limits_{S \in X_k} [p_w(S) \geqslant \theta]$ [12]. Подставляя эти оценки в плотности $\pi_k(S)$, затем эти плотности – в формулу (1), снова получим линейную дискриминантную функцию, но с другими коэффициентами $\gamma_{kw}$:

$$b_k(S) = \sum_{w \in W} \gamma_{kw} [p_w(S) \geqslant \theta], \quad \gamma_{kw} = \ln \frac{B_w(X_k, \theta)(1 - B_w(X_0, \theta))}{B_w(X_0, \theta)(1 - B_w(X_k, \theta))}.$$

Данная модель называется *синдромным алгоритмом* [12]. Она позволяет осуществлять отбор признаков, выделяя короткие диагностические эталоны каждого заболевания. Именно эти триграммы должны получать наибольшие по модулю веса в линейном классификаторе. Использование остальных «шумовых» триграмм может приводить к снижению информативности дискриминантной функции и падению качества классификации. Чтобы этого не происходило, предлагается ранжировать триграммы по убыванию встречаемости $B_w(X_k, \theta)$ в классе $y_k$, и учитывать первые $H$ триграмм. Число $H$ является параметром метода и подбирается в эксперименте. Полученные наборы информативных $H$ триграмм будут образовывать диагностические эталоны заболевания.

## 3.2 Вероятностная тематическая модель двухклассовой классификации

Пусть $D$ – конечное множество (коллекция) текстовых документов, $T$ – конечное множество тем. $M$ – конечное множество модальностей. Каждой модальности $m \in M$ соответствует словарь – конечное множество токенов $V^m$. Каждый документ $d \in D$ представляет собой последовательность слов $v_1, \ldots, v_{n_d}$ из $V = \bigcup\limits_{m=1}^{M} V^m$ , где $n_d$ – длина документа. Принимая «гипотезу мешка слов», будем считать, что последовательность токенов не важна, и учитывать только число вхождений $n_{dv}$ токена $v$ в документ $d$.

Вероятностная тематическая модель описывает условную вероятность появле-

ния токена $v$ в документе $d \in D$ как вероятностную смесь распределений $\phi_{vt} = p(v|t)$ токенов в темах $t \in T$ с коэффициентами $\theta_{td} = p(t|d)$, зависящими от документов:

$$p(v|d) = \sum_{t \in T} \phi_{vt} \theta_{td}, \quad v \in V^m, \quad d \in D.$$

Матрицы $\Phi = (\phi_{vt})_{V \times T}$ и $\Theta = (\theta_{td})_{T \times D}$ будем использовать для обозначения параметров тематической модели.

В. М. Успенский впервые ввел понятие диагностических эталонов заболевания. В его исследованиях количество таких наборов для каждого заболевания было заранее фиксированной величиной. В. М. Успенский предполагал, что длина выделяемых диагностических эталонов должна быть небольшой. Для удобства дальнейшего изложения данное предположение формализуем как возможность получения интерпретируемой модели. В данной работе предлагается осуществить автоматическое выделение диагностических эталонов каждого заболевания. Для получения интерпретируемой модели необходим поиск коротких диагностических эталонов. Для этого будем вводить дополнительные ограничения.

Вероятностный латентный семантический анализ (PLSA) и латентное размещение Дирихле (LDA) используют единственную модальность терминов (как правило, отдельных слов). Ставится задача максимизации логарифма правдоподобия модели $p(v|d)$ при ограничениях неотрицательности и нормированности столбцов матриц $\Phi$ и $\Theta$ [13]. К сожалению, модели PLSA и LDA не позволяют учитывать различные дополнительные требования и ограничения, возникающие в прикладных задачах. Для построения таких расширений в [14] была предложена теория аддитивной регуляризации тематических моделей (ARTM), а в [15] показано, как использовать механизм модальностей для построения тематических моделей классификации.

В аддитивной регуляризации тематических моделей (ARTM) критерий логарифма правдоподобия вводится для каждой модальности и максимизируется их взвешенная сумма. Дополнительные ограничения накладываются при помощи критериев-регуляризаторов $R_i$:

$$\sum_{m \in M} \frac{\tau_m}{n_m} \sum_{d \in D} \sum_{v \in V^m} n_{dv} \ln \sum_{t \in T} \phi_{vt} \theta_{td} + \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \ \to \max_{\Phi, \Theta}$$

где $n_m = \sum\limits_{d \in D} \sum\limits_{v \in V^m} n_{dv}$ – нормировочный множитель для балансировки модальностей.

Задача оптимизации решается с помощью регуляризованного EM-алгоритма [15].

*Регуляризатор разреживания* $\phi_{vt}$, $v \in V^m$ вводит в модель требование, чтобы каждая тема состояла из небольшого числа слов словаря $V^m$, вероятности остальных слов в распределении $\phi_{vt}$ равны 0.

$$R(\Phi) = -\tau_1 \sum_{t \in T} \sum_{v \in V^m} \ln \phi_{vt} \ \to \max_{\Phi,\Theta};$$

Чем больше коэффициент регуляризации $\tau_1$, тем больше вероятностей $\phi_{vt}$ обращается в 0 на каждой итерации.

*Регуляризатор декоррелирования* $\phi_{vt}$, $v \in V^m$ минимизирует ковариации между вектор-столбцами матрицы $\phi_{vt}$, что способствует повышению различимости тем.

$$R(\Phi) = -\tau_2 \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{v \in V^m} \phi_{vt}\phi_{vs} \ \to \max_{\Phi,\Theta};$$

Декоррелирование приводит также к разреживанию тем и к более четкому выделению ядер тем, состоящих из слов $v$ с сильно доминирующей вероятностью $p(t|v)$.
На практике два описанных регуляризатора показывают хорошие результаты, работая совместно [16].

В данной работе рассматривается двумодальная регуляризованная тематическая модель двухклассовой классификации. Примерами модальностей являются триграммы и метки классов. Обозначим через $C$ – словарь меток классов ($|C| = 2$), через $W$ – словарь триграмм ($|W| = 216$). Тематическая модель классификации описывает распределение меток классов в документах $p(c|d)$ через распределения $p(c|t)$ и $p(t|d)$. Для меток классов постулируется гипотеза условной независимости $p(c|t,d) = p(c|t)$, означающая, что для классификации документа достаточно знать его тематику:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d), \quad c \in C, \quad d \in D$$

Все описанные регуляризаторы вводятся только для словаря триграмм $W$. Высокая разреженность $\phi_{wt}$, $w \in W$ способствует повышению интерпретируемости модели.

# 4 Эксперименты

В данном разделе приводятся результаты обучения тематической модели, наивного байесовского классификатора и синдромного алгоритма. Для настройки параметров и выбора лучшей версии каждого из алгоритмов используется 10×10-кратная кросс-валидация. Для избежания получения оптимистично завышенных оценок качества, работа итоговых моделей оценивается по отложенной выборке.

## 4.1 Оценки качества диагностики

Для оценивания качества модели возникает несколько естественных мер – чувствительность и специфичнсоть. Чувствительность – это доля правильно определенных моделью больных. Специфичность – доля правильно определенных моделью здоровых. Однако, в зависимости от выбранного порога классификации, соотношение чувствительности-специфичности может быть разным, поэтому для сравнения моделей будем считать следующую меру качества:
AUC (Area Under Curve) – площадь под ROC-кривой в координатах чувствительность-специфичность, которая получается при варьировании порога классификации:

$$AUC = \frac{1}{|X_0||X_k|} \sum_{d \in X_0} \sum_{d' \in X_k} [p(c|d) > p(c|d')].$$

Однако, данная мера не позволяет оценивать сами значения вероятностей, которые являются очень важной характеристикой модели. Для этого введем дополнительную меру качества диагностики – меру LogLoss (Logistic Loss), позволяющую сравнивать вероятностные модели:

$$LogLoss = -[c = k] \log p(c|d) - [c = 0] \log(1 - p(c|d)).$$

Для оценивания интерпретируемости тематической модели воспользуемся мерой разреженности, равной доле нулевых значений в следующих матрицах:

- $\mathbf{p}(\mathbf{w}|\mathbf{t}), \mathbf{w} \in \mathbf{W}$.

  Эксперименты показывают возможность разреживания до 80 – 90% без существенной потери качества диагностики. Таким образом, тематическая модель позволяет выделять короткие диагностические эталоны.

- $\mathbf{p(t|c)}, \mathbf{c} \in \mathbf{C}$

$$p(t|c) = \frac{p(c|t)p(t)}{p(c)}, \quad c \in C, \quad t \in T;$$

$$p(t) = \sum_{d \in X_0 \cup X_k} p(t|d)p(d), \quad p(d) = \frac{1}{|X_0 \cup X_k|}, \quad p(c = k) = \frac{|X_k|}{|X_0 \cup X_k|}.$$

Разреженность распределения $\mathbf{p(t|c)}$ достигает 50%, что позволяет говорить о выделении непересекающихся наборов тем в каждом классе $c$.

## 4.2 Вероятностная тематическая модель двухклассовой классификации

В данном разделе предлагается рассмотреть ряд последовательных экспериментов с тематической моделью. В каждом следующем эксперименте добавляются новые эвристики, улучшающие качество и интерпретируемость модели. Результаты работы тематической модели сравниваются с основными конкурирующими методами: наивным байесовским классификатором (NB) и синдромным алгоритмом (SA). Во всех экспериментах модель SA была разрежена до 80%, модель NB – не разрежена. Выход обеих моделей калибровался при помощи калибровки Платта [17]. Эксперименты с тематической моделью проводились при помощи библиотеки BigARTM. Болезнь «Холецистит хронический» (XX) рассматривается в качестве демонстрационной. Поведение каждой модели демонстрируется на графиках и в таблицах. Для мер качества AUC и LogLoss дополнительно показывается доверительный интервал 95%.

### 4.2.1 PLSA

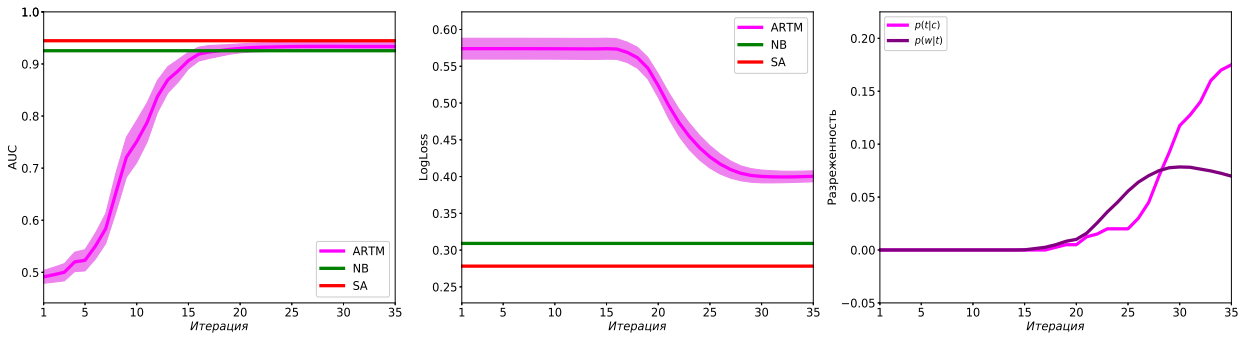



Рис. 1: Зависимость AUC, LogLoss и разреженности от числа итераций для модели PLSA, сравнение с NB и SA.

В данном эксперименте предлагается сравнить модель PLSA с моделями SA и NB, показавшими высокие результаты в работе [12]. Проверяется основная гипотеза о возможности получения высокого качества простой унимодальной тематической модели с $|T| = 2$. Матрица $\Phi$ инициализируется случайным распределением. Матрица $\Theta$ здесь и далее инициализируется равномерным распределением.

На рисунке 1 показаны зависимости качества AUC, LogLoss и разреженности от итерации EM-алгоритма. Горизонтальными линиями обозначено качество основных конкурентов – модели SA и NB. Лучшей с точки зрения AUC и LogLoss оказалась модель SA. Модель NB показывает более низкое качество AUC по сравению с моделью PLSA. Последняя демонстрирует самый слабый результат с точки зрения меры LogLoss.

Несмотря на то, что разреженность распределения $\mathbf{p}(\mathbf{t}|\mathbf{c})$ на последних 5 итерациях начала немного расти, модель становилась все менее интерпретируемой, так как наблюдалось уменьшение разреженности распределения $\mathbf{p}(\mathbf{w}|\mathbf{t})$. В результате, показав высокое качество в смысле меры AUC, модель PLSA не смогла выделить короткие диагностические эталоны, а также сформировать непересекающиеся наборы тем в каждом классе.

### 4.2.2 Подбор весов модальностей

Основной целью эксперимента является формирование непересекающихся наборов тем, описывающих классы больных и здоровых людей. Для проведения эксперимента была взята мультимодальная модель ARTM. Она позволяет выбрать вес каждой модальности. Чем больше вес выбранной модальности, тем лучше будет она описываться. В данной работе рассматривается двумодальная тематическая модель, поэтому достаточно задать вес только для модальности метки классов. Обозначим его $\tau$. На рисунке 2 показаны зависимости мер качества AUC, LogLoss и разреженности от числа итераций при различных значениях параметра $\tau$. Для каждого значения параметра $\tau$ проводилось 25 итераций EM-алгоритма. В результате было выбрано значение $\tau = 10^3$. Жирной кривой выделена наилучшая траектория. На рисунке 3 показаны результаты итоговой модели. Наблюдается улучшение качества в смысле меры AUC (результат ARTM совпал с SA), ухудшение в смысле меры LogLoss. Модель по-прежнему слабо интерпретируема (полное отсутствие разреженности распределения $p(w|t), w \in W$). Удалось добиться нужной степени разреженности матрицы

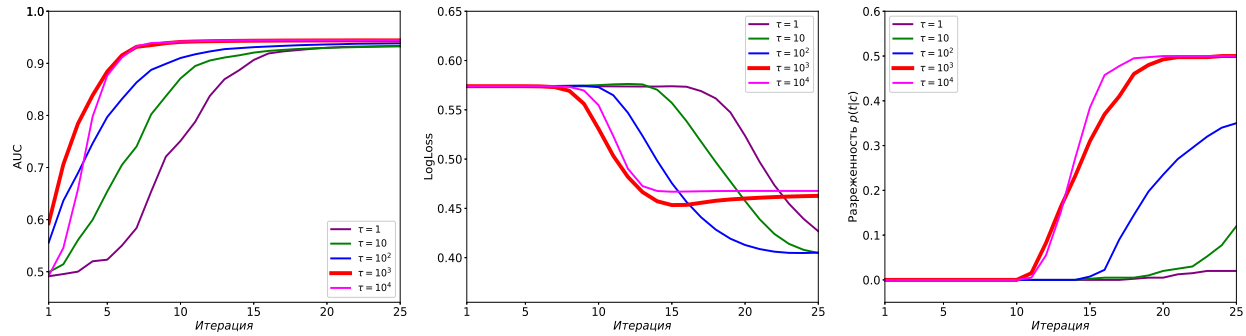



Рис. 2: Зависимость AUC, LogLoss и разреженности от числа итераций и параметра $\tau$, модель ARTM.

$p(t|c)$. Выделились наборы тем, характеризующие каждый класс.

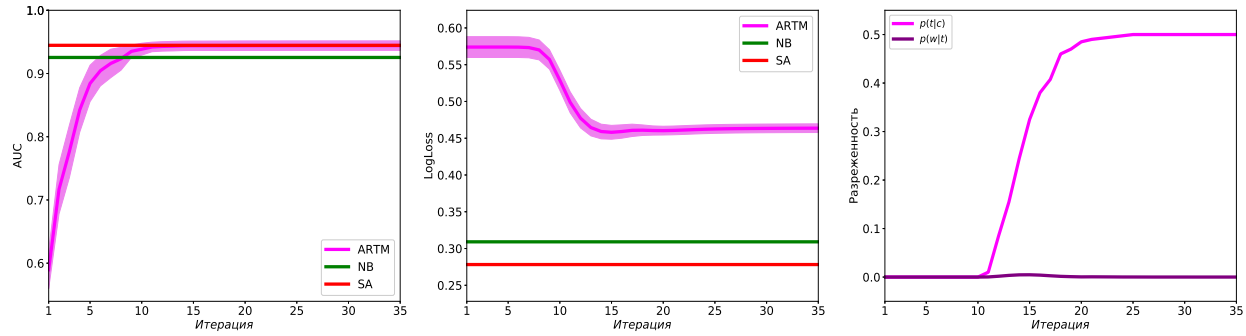



Рис. 3: Зависимость AUC, LogLoss и разреженности от числа итераций для мультимодальной модели ARTM, сравнение с NB и SA.

### 4.2.3 Подбор инициализации методами кластеризации

В данном эксперименте предлагается опробовать различные эвристические способы инициализации матрицы Φ тематической модели. Эта задача является крайне важной для проведения будущих исследований многоклассовой классификации в задачах медицинской диагностики. Поэтому результаты экспериментов будут представлены сразу для всех болезней.

Тематическая модель осуществляет мягкую кластеризацию, а значит, логичнее всего ее инициализировать структурно схожими моделями. Рассмотрим некоторые из них.

Задачи кластеризации отличаются от классификации тем, что в них не задаются ответы и требуется разбить выборку на подмножества (кластеры) так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно

отличались. В данной работе рассматриваются методы кластеризации, в которых необходимо задавать функцию расстояния на множестве объектов. В качестве такой функции будет рассмотрено евклидово расстояние. Количество кластеров будет равным выбранному числу тем модели. Найденные центры кластеров будут формировать столбцы матрицы Ф. Везде, где не указано иное, кластеризация будет проводиться по документам.

Одним из популярных методов кластеризации является модель *KMeans*, фактически реализующая жесткий вариант EM-алгоритма: на E-шаге осуществляется пересчет центров кластеров, на M-шаге пересчитываются кластеры по степени близости к центру.

Следующий метод является упрощением модели *KMeans*. Поэтому он называется наивным *KMeans* (*NaiveKMeans*). Данная модель, в отличие от обычного *KMeans*, центрами кластеров назначает объекты, максимально удаленные ото всех объектов выборки. Принимая наивное предположение, что таким образом мы сможем найти кластеры, лучше всего описывающие распределение всей выборки.

Его обобщением является метод, основанный на алгоритме выделения $\epsilon$-кластерной структуры (*Spread*) [18]. Суть метода состоит в формировании некоторого множества объектов, позволяющих наиболее полно описать всю выборку. Метод является итерационным, поэтому требует введения критерия останова. В работе [18] останов осуществлялся в тот момент, когда каждый объект выборки принадлежал некоторому шару радиуса $\epsilon$. В данной работе останов будем осуществлять, когда наберем нужное количество центров. На первом шаге алгоритма ищутся два объекта, максимально удаленные друг от друга. Затем они добавляются в некоторое множество $\Omega$. Далее для каждого объекта множества $\Omega$ вычисляются два ближайших объекта, не принадлежащие множеству $\Omega$. Новым центром объявляется тот объект, расстояние от которого до ближайшего его соседа множества $\Omega$ является максимальным.

Следующий метод основан на понятии коэффициента гармонии (КГ) кодограммы, предложенного В. М. Успенским (*HarmonyBaskets*). Коэффициент гармонии представлят собой отношение числа A, B, E, F к числу C, D в кодограмме. В своих исследованиях В. М. Успенский фиксировал число кластеров, одинаковое для всех болезней. Объекты с близкими значениями КГ далее формировали один кластер. Данный метод использовался для поиска диагностических эталонов каждого заболевания. В настоящей работе ставится задача автоматического подбора числа

кластеров для каждой болезни.

Наконец, логичным продолжением кластеризации документов является кластеризация триграмм. Этот метод назовем *KMeansGram3*. Такой подход к инициализации тематических моделей показал высокое качество в работе [19]. Алгоритм кластеризации практически не меняется – только кластеризуются не документы, а триграммы, описывающие документы. Выделенные центры кластеров формируют столбцы матрицы Φ, отсутствующие триграммы заполняются равномерным распределением.

Основной гипотезой данного эксперимента является возможность улучшения качества диагностики мультимодальной модели ARTM путем подбора инициализации. В своих исследованиях В. М. Успенский предполагал, что по 4 диагностических эталона для класса больных и здоровых людей вполне достаточно. Поэтому, для демонстрации результатов данного эксперимента было зафиксировано число тем модели $|T| = 8$. При этом 2 темы были выделены для класса здоровых людей, 6 – для класса больных. Значение параметра $\tau$ было взято из предыдущего эксперимента ($\tau = 10^3$).

В таблице 2 представлены значения меры LogLoss мультимодальной модели ARTM для каждой болезни в зависимости от способа инициализации. В среднем по всем болезням лучшей инициализацией оказался алгоритм В. М. Успенского (метод HarmonyBaskets). Аналогичная таблица для меры качества AUC не представлена, так как в ходе экспериментов наблюдалось слабое улучшение качества диагностики в зависимости от способа инициализации.

На рисунке 4 видно улучшение качества LogLoss по сравнению с предыдущим экспериментом, разреженность распределения $p(t|c)$ осталась на прежнем уровне. Побочным эффектом текущего эксперимента является увеличение разреженности распределения $p(w|t)$. Выделяются более короткие диагностические эталоны по сравнению с предыдущими экспериментами.

### 4.2.4 Подбор регуляризации

Основной целью данного эксперимента является выделение коротких диагностических эталонов. Предлагается включить два регуляризатора: декорреляцию и разреживание распределения $p(w|t)$. Регуляризаторы добавлялись в модель в указанном порядке один за другим. При добавлении каждого регуляризатора его коэффициент регуляризации выбирался из заданной сетки значений в соответствии с критериями качества AUC и LogLoss. Первые 10 итераций включалась только декорреляция,

| | KMeans | Random | Naive KMeans | Spread | KMeans Gram3 | Harmony Baskets |
|---|---|---|---|---|---|---|
| ММ | **0.42** | 0.46 | **0.42** | 0.43 | **0.42** | **0.42** |
| ХХ | 0.39 | 0.42 | 0.4 | 0.4 | **0.38** | **0.38** |
| ВСД | **0.46** | 0.48 | 0.48 | **0.46** | 0.47 | **0.46** |
| ГД | 0.49 | 0.52 | 0.5 | 0.49 | 0.49 | **0.48** |
| ЯБ | **0.39** | 0.41 | **0.39** | 0.4 | **0.39** | **0.39** |
| ГБ | 0.22 | 0.22 | 0.22 | 0.24 | 0.23 | **0.21** |
| ИБС | 0.27 | 0.26 | **0.24** | 0.25 | 0.27 | 0.26 |
| СД | **0.32** | 0.34 | **0.32** | **0.32** | 0.33 | **0.32** |
| МКБ | 0.4 | 0.44 | 0.41 | 0.4 | 0.4 | **0.39** |
| ЖКБ | 0.44 | 0.49 | 0.45 | 0.44 | 0.44 | **0.43** |
| УЩ | 0.41 | 0.42 | 0.4 | 0.41 | 0.4 | **0.39** |
| ХГ | 0.39 | 0.41 | 0.39 | 0.4 | **0.37** | **0.37** |
| ДЖВП | **0.37** | 0.41 | 0.38 | 0.39 | 0.38 | **0.37** |
| РО | **0.43** | 0.46 | 0.46 | 0.45 | 0.44 | **0.43** |
| ДГПЖ | 0.48 | 0.52 | 0.51 | **0.47** | 0.48 | 0.48 |
| АХ | **0.52** | 0.57 | 0.58 | **0.52** | 0.53 | 0.54 |
| ЖДА | **0.56** | 0.61 | 0.58 | 0.57 | 0.59 | 0.61 |
| НГБК | **0.49** | 0.53 | 0.5 | 0.5 | **0.49** | 0.52 |
| ТБЦ | **0.44** | 0.49 | 0.45 | 0.45 | 0.45 | **0.44** |
| ОП | 0.43 | 0.43 | 0.43 | 0.44 | **0.39** | 0.43 |

Таблица 2: Значения LogLoss на тестовых выборках кросс-валидации при использовании разных методов инициализации, мультимодальная модель ARTM.

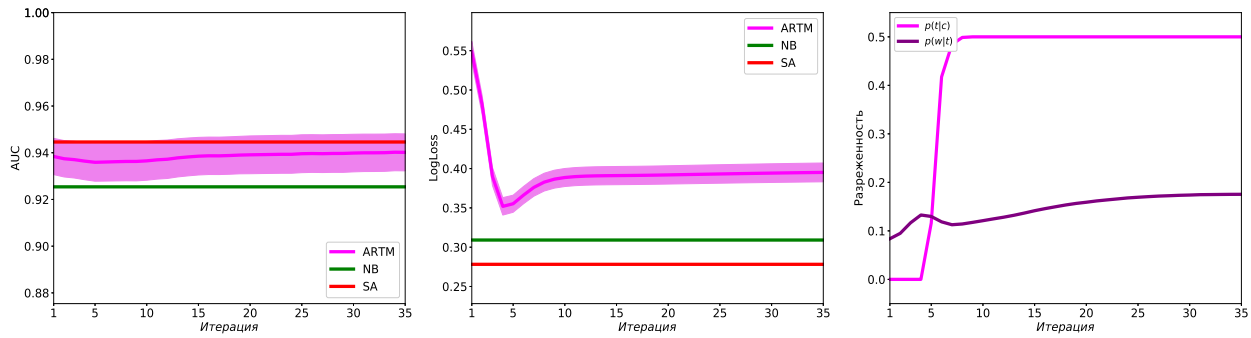



Рис. 4: Зависимость AUC, LogLoss и разреженности от числа итераций для модели ARTM, инициализированной методом В. М. Успенского, сравнение с NB и SA.

затем на 11 итерации добавлялось разреживание, сила которого постепенно увеличивалась от итерации. Из всех значений выбиралось то, при котором улучшался хотя бы один из критериев без существенного ухудшения другого [20]. На рисунке 5 показаны зависимости AUC, LogLoss и разреженности от числа итераций при различных значениях коэффициентов регуляризации. В результате была выбрана совокупность коэффициентов регуляризации $\tau_1 = 10$, $\tau_2 = 1$. Жирной кривой выделена наилуч-

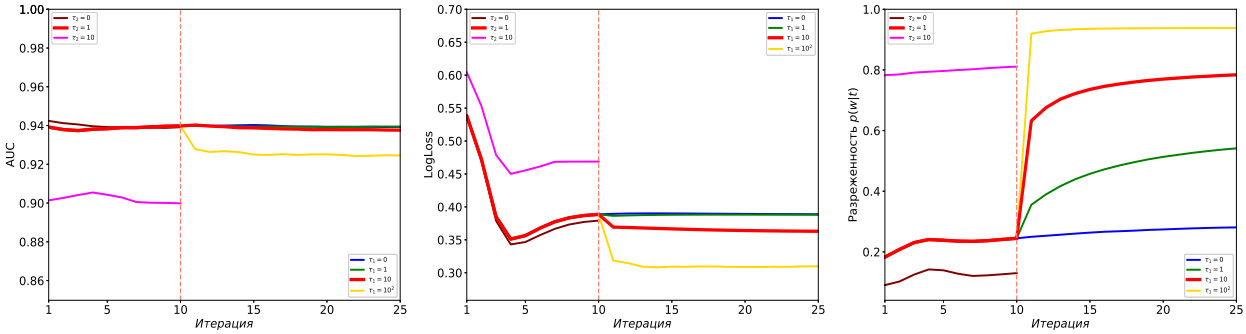



Рис. 5: Зависимость AUC, LogLoss и разреженности от числа итераций и коэффициентов регуляризации, модель ARTM.

шая траектория регуляризации. На рисунке 6 показаны результаты итоговой моде-

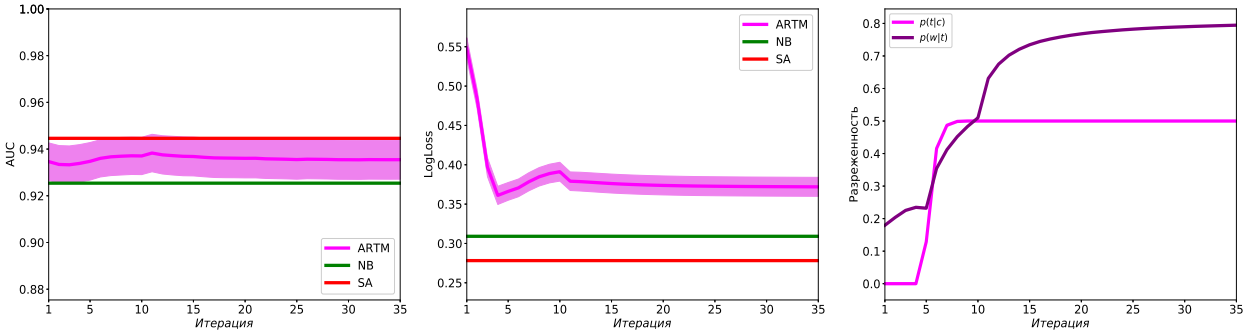
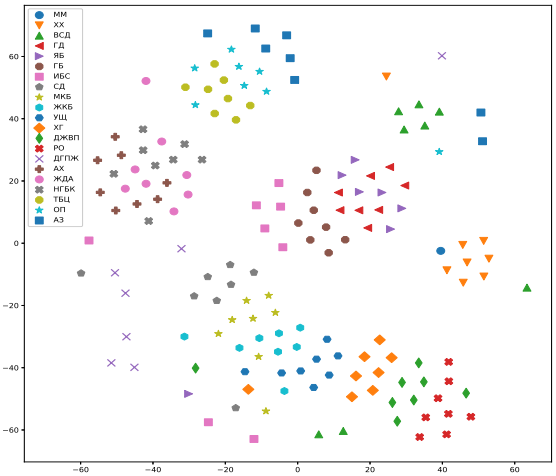



Рис. 6: Зависимость AUC, LogLoss и разреженности от числа итераций для разреженной модели ARTM, сравнение с NB и SA.

ли. Декоррелирующий регуляризатор обладает побочным эффектом разреживания, поэтому уже на первой итерации распределение $p(w|t)$ было разрежено до 80%. Добавление разреживающего регуляризатора, начиная с 15 итерации позволило найти короткие диагностические эталоны (в каждом классе порядка 40 информативных триграмм).

Данный результат является важным с точки зрения основной гипотезы настоящей работы. Поэтому продемонстрируем его сразу для всех болезней. На рисунке 7 справа представлены самые длинные диагностические эталоны, описывающие одну тему. Слева визуализировано распределение диагностических эталонов для каждой болезни. Понижение размерности проводилось при помощи метода t-SNE с использованием евклидовой меры расстояний между объектами. Выделяются кластеры, описывающие похожие болезни. Так, например, желчнокаменная (ЖКБ), мочекаменная болезни (МКБ) часто встречаются у людей, больных дискинезией желчно-

выводящих путей (ДЖВП). Ишемическая болезнь сердца (ИБС) и гипертоническая болезнь (ГБ) также формируют единый кластер.




| | |
|---|---|
| ММ | DFC, CAD, ADF, FCC, ADD, CDF DDA |
| ХХ | FAE, DFE, FEF, ACC, AFA, FEC, AFB |
| ВСД | BAF, BCB, AED, DDB, CEF, CFD, EDB |
| ГД | FFC, FCD, FED, AEC, FBC, EAB, EAE |
| ЯБ | CCD, DEA, DAC, FBC, EDD, EAB, CAD |
| ГБ | FBF, DDB, BAD, FDF, DAC, DDE, BCA |
| ИБС | CDD, DDC, CCD, DFC, DCC, CFD, DCD |
| СД | BCF, DAB, BFD, FDB, ACC, FDF, ACE |
| МКБ | DFB, AED, DED, CED, EAC, BBB, FBC |
| ЖКБ | DAF, AAB, FFE, FAB, ABF, ACB, BFD |
| УЩ | BAD, FFE, CFF, EDF, BAC, CFC, ACD |
| ХГ | DDB, BAD, FAC, ECB, EAB, DAC, DDE |
| ДЖВП | EFC, ADE, DCE, DDC, DBC, CCC, BBA |
| РО | BDD, CDE, DCF, DCD, BDF, CBF, ECC |
| ДГПЖ | CAE, BCE, BEF, CDE, BEB, FCD, CFE |
| АХ | EAD, FCD, FFC, FFF, ACA, DAA, FDD |
| ЖДА | ACA, CAC, EAC, DAD, BAB, BBF, ADA |
| НГБК | EEF, FBA, FBC, EAB, FBE, BBE, EDB |
| ТБЦ | CDC, DAC, FDE, AFB, FDC, FFB, CBC |
| ОП | CCD, ADB, EBF, DDF, DBC, FCB, DFB |
| АЗ | EDB, BBC, EEB, DBC, FBC, BCB, BBB |

Рис. 7: Диагностические эталоны каждой болезни, описывающие одну тему.

## 4.3 Сравнение методов

В таблицах рисунка 8 приведены значения AUC и LogLoss соответственно на тестовых выборках всех представленных болезней для следующих моделей: *PLSA*, *MultiARTM* – двумодальная ARTM, *SparseARTM* – регуляризованная ARTM, *SmartARTM* – ARTM с автоматическим подбором всех параметров, *NB* – наивный байесовский классификатор и *SA* – синдромный алгоритм с подобранным значением параметра $H$. Самые высокие результаты показала модель SA, которая вместе с моделью NB практически не склонна к переобучению. Модель SmartARTM подбирала параметры при помощи жадного алгоритма. К основным параметрам, рассмотренных в экспериментах добавлялось число тем $|T|$. Ставилась гипотеза о возможности независимого подбора числа тема $|T|$ и способа инициализации. На основании результатов, представленных для всех болезней можно утверждать, что автоматический подбор совокупности всех параметров практически не ухудшает качество классификации. В большинстве своем более высокое качество демонстрировали модели с 2, 4 и 8 темами. То есть тематическая модель по сложности была близка модели NB, поэтому и показала сравнимые результаты.

| | PLSA | Multi ARTM | Sparse ARTM | Smart ARTM | NB | SA |
|---|---|---|---|---|---|---|
| ММ | 0.89 | 0.89 | 0.87 | 0.89 | 0.88 | **0.92** |
| ХХ | 0.9 | 0.9 | 0.9 | 0.89 | 0.9 | **0.92** |
| ВСД | 0.77 | **0.83** | 0.81 | **0.83** | 0.82 | **0.83** |
| ГД | 0.89 | 0.92 | 0.91 | 0.91 | 0.91 | **0.96** |
| ЯБ | 0.9 | 0.91 | 0.9 | 0.9 | 0.9 | **0.95** |
| ГБ | 0.86 | 0.92 | 0.9 | 0.91 | 0.91 | **0.95** |
| ИБС | 0.92 | 0.94 | 0.93 | 0.93 | 0.93 | **0.97** |
| СД | 0.93 | **0.94** | **0.94** | 0.91 | **0.94** | **0.94** |
| МКБ | 0.9 | 0.91 | 0.89 | 0.9 | 0.9 | **0.92** |
| ЖКБ | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | **0.98** |
| УЩ | 0.91 | 0.92 | 0.91 | 0.92 | 0.92 | **0.93** |
| ХГ | 0.92 | **0.93** | 0.92 | 0.92 | 0.92 | **0.93** |
| ДЖВП | 0.9 | 0.91 | 0.9 | 0.88 | 0.9 | **0.92** |
| РО | 0.92 | **0.94** | 0.93 | 0.92 | 0.93 | **0.94** |
| ДГПЖ | 0.94 | **0.95** | **0.95** | **0.95** | **0.95** | **0.95** |
| АХ | 0.83 | **0.86** | **0.86** | 0.85 | **0.86** | **0.86** |
| ЖДА | 0.83 | 0.88 | 0.87 | 0.87 | 0.88 | **0.89** |
| НГБК | 0.93 | 0.95 | 0.95 | 0.95 | 0.95 | **0.97** |
| ТБЦ | 0.91 | 0.93 | 0.91 | 0.91 | 0.93 | **0.98** |
| ОП | 0.97 | **0.98** | **0.98** | **0.98** | **0.98** | **0.98** |

| PLSA | Multi ARTM | Sparse ARTM | Smart ARTM | NB | SA |
|---|---|---|---|---|---|
| 0.43 | 0.53 | 0.42 | 0.4 | 0.38 | **0.34** |
| 0.47 | 0.51 | 0.42 | 0.42 | 0.39 | **0.35** |
| 0.56 | 0.59 | 0.51 | **0.48** | 0.49 | **0.48** |
| 0.47 | 0.52 | 0.45 | 0.43 | 0.36 | **0.25** |
| 0.42 | 0.51 | 0.44 | 0.42 | 0.37 | **0.3** |
| 0.44 | 0.47 | 0.28 | 0.29 | 0.25 | **0.2** |
| 0.54 | 0.44 | 0.31 | 0.31 | 0.27 | **0.19** |
| 0.59 | 0.46 | 0.36 | 0.39 | 0.32 | **0.28** |
| 0.44 | 0.51 | 0.43 | 0.43 | 0.37 | **0.35** |
| 0.41 | 0.46 | 0.36 | 0.35 | 0.28 | **0.21** |
| 0.62 | 0.49 | 0.4 | 0.46 | 0.36 | **0.33** |
| 0.62 | 0.49 | 0.38 | 0.38 | **0.32** | **0.32** |
| 0.48 | 0.52 | 0.43 | 0.41 | 0.37 | **0.33** |
| 0.41 | 0.48 | 0.37 | 0.41 | 0.33 | **0.3** |
| 0.42 | 0.47 | 0.38 | 0.39 | 0.31 | **0.29** |
| 0.56 | 0.56 | 0.51 | 0.5 | **0.47** | **0.47** |
| 0.62 | 0.55 | 0.48 | **0.45** | **0.45** | **0.45** |
| 0.46 | 0.48 | 0.39 | 0.38 | 0.28 | **0.19** |
| 0.51 | 0.5 | 0.43 | 0.46 | 0.33 | **0.19** |
| 0.37 | 0.4 | 0.3 | 0.31 | 0.21 | **0.18** |

Рис. 8: Значения AUC и LogLoss при использовании разных моделей классификации.

# 5 Заключение

Данная работа иллюстрирует применение нового подхода в информационном анализе электрокардиосигналов – использование регуляризованной тематической модели классификации. Подтвердилась основная гипотеза о возможности автоматического выделения диагностических эталонов для всех рассмотренных заболеваний. Продемонстрированные результаты позволяют говорить о выделении коротких даигностических эталонов практически без потери качества диагностики. Предложен алгоритм, обеспечивающий автоматический подбор параметров тематической модели, оптимизирующий совокупность критериев качества для любого из представленных заболеваний. Высокое качество классификации при обязательном требовании интерпретируемости позволяет использовать тематическую модель для осуществления качественной диагностики.

# 6 Список литературы

[1] Баевский Р. М., Иванов Г. Г. Вариабельность сердечного ритма: теоретические аспекты и возможности клинического применения. //Ультразвуковая и функциональная диагностика. – 2001. – no. 3. – С. 108–127.

[2] Баевский Р. М., Иванов Г. Г., Чирейкин Л. В., Гаврилушкин А. П., Довгалевский П. Я., Кукушкин Ю. А., Миронова Т. Ф. и др. Анализ вариабельности сердечного ритма при использовании различных электрокардиографических систем (методические рекомендации). //Вестник аритмологии. – 2001. – no 24. – С. 65–87.

[3] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. // Экономика и информатика. – 2008. – 116 с.

[4] Uspenskiy V. M. Information Function of the Heart. Biophysical substantiation of technical requirements for electrocardioblock registration and measurement of electrocardiosignals parameters acceptable for information analysis to diagnose internal diseases. //In: Joint International IMEKO TC1+TC7+TC13 Symposium. Jena, Germany. – 2011.

[5] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, USA. – 1999. – Pp. 50–57.

[6] Uspenskiy V. M. Information Function of the Heart. A Measurement Model. // In: Measurement 2011: 8-th International Conference. Smolenice, Slovakia. – 2011. – Pp. 383–386.

[7] Uspenskiy V. M. Diagnostic System Based on the Information Analysis of Electrocardiogram. //In: Proceedings of MECO 2012. Advances and Challenges in Embedded Computing. Bar, Montenegro. – 2012. – Pp. 74–76.

[8] J. Kurths, A. Voss, P. Saparin, A. Witt, H.J. Kleiner, and N.Wessel Quantitative analysis of heart rate variability Chaos. – 1995. – no 1. – Pp. 88–94.

[9] Albert C. C. Yang, Shu-Shya Hseu, Huey-Wen Yien, Ary L. Goldberger and C. K. Peng Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics Physical review letters. – 2003. – no 10.

[10] Camillo Cammarota, Enrico Rogora Time reversal, symbolic series and irreversibility of human heartbeat Chaos, Solitons & Fractals. – 2007. – no 5. – Pp. 1649–1654.

[11] U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths and N. Wessel Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics Computers in Biology and Medicine. – 2012. – no 3. – Pp. 319–327.

[12] Целых В. Р. Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. // Магистерская диссертация. МФТИ. – 2015.

[13] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. – New York, USA. – 1999. – Pp. 50–57.

[14] Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. – 2014. – Т. 455. – no. 3. – Pp. 268–271.

[15] Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A., Yanina A. O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. Melbourne, Australia. New York, USA. – 2015. – Pp. 29–37.

[16] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications. – 2015. – no. 1. – Pp. 303–323.

[17] Hsuan-Tien Lin, Chih-Jen Lin, Ruby C. Weng 2000. A Note on Platt's Probabilistic Outputs for Support Vector Machines.

[18] Вальков А. С. Субквадратичные алгоритмы метрического анализа данных. Диссертация на соискание ученой степени кандидата физико-математических наук. //Российская академия наук Вычислительный центр им. А.А. Дородницына РАН. –

2005.

[19] Ивашковский И. А. Методы инициализации в вероятностном тематическом моделировании. // Выпускная квалификационная работа. МФТИ. – 2016.

[20] Янина А. О., Воронцов К. В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных. – Т.2. – no 2. – 2016.