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

2017

Содержание

Аннотация

Технология информационного анализа электрокардиосигналов, основанная на теории информационной функции сердца, уже более 15 лет применяется во врачебной практике для определения рисков наиболее опасных и распространенных заболеваний внутренних органов. За это время была накоплена выборка, содержащая более 20 тысяч обследований как больных различными заболеваниями, так и здоровых людей. На основании этих результатов в данной работе осуществляется построение конкурентноспособной и максимально интерпретируемой тематической модели классификации. Проводится сравнение с существующими линейными моделями, показавшими высокое качество для всех заболеваний [1].

1 Введение

Прежде, чем приступать к пониманию природы ЭКГ-сигналов, поступающих с электрокардиографа, необходимо предложить такой способ их обработки, который бы, во-первых, позволил сократить в десятки раз объем информации и, во-вторых, облегчил бы ее восприятие для возможности дальнейшей работы. Один из таких способов обработки был предложен более 15 лет назад профессором В.М.Успенским при разработке метода информационного анализа электрокардиосигналов, основанного на предположении о том, что амплитуды и интервалы кардиоциклов несут в себе информацию не только о состоянии сердца, но и о состоянии всего организма человека.

2 Предварительная обработка ЭКГ-сигнала

Технология информационного анализа электрокардиосигналов включает три этапа их предварительной обработки: вычисление интервалов и амплитуд, дискретизацию и векторизацию [2].

2.1 Вычисление интервалов и амплитуд

Электрокардиограмма (ЭКГ) представляет собой квазипериодический сигнал, периоды которого называются кардиоциклами, каждый из которых описывается тройкой

(R_n, T_n, α_n), где R_n – амплитуда n-го кардиоцикла, T_n – интервал n-го кардиоцикла, α_n – аналог фазового угла в гармонических сигналах, определяемый как арктангенс отношения амплитуды к интервалу: $n = \text{arctg}\,\frac{R_n}{T_n}$. Статистическая ценность данного подхода была доказана в работе [1].

2.2 Дискретизация

Согласно методу информационного анализа электрокардиосигналов, диагностическую ценность имеют не столько измеряемые величины амплитуд R_n, интервалов T_n и углов α_n, сколько знаки их приращений в последовательных кардиоциклах. Возможны только 6 сочетаний увеличений и уменьшений этих трех величин, которые кодируются первыми буквами 6-символьного алфавита: $\{A, B, C, D, E, F\}$:

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

2.3 Векторизация

Последовательность символов $S = (s_n)_{n=1}^N$ называется кодограммой. Каждые 3 последовательных символа (s_{n-1}, s_n, s_{n+1}) образовывают триграмму. Существуют наборы триграмм, совместная встречаемость которых говорит о наличии в организме определенного заболевания. Согласно [1] значительная доля триграмм может быть отброшена без существенной потери качества диагностики. Как будет показано, порядка 20% информативных триграмм бывает вполне достаточно. Каждую триграмму можно интерпретировать как слово из словаря W^{gram3}, содержащего $6^3 = 216$ слов. Набор триграмм, описывающий отдельную кодограмму, интерпретируется как документ, а в сумме закодированные электрокардиограммы являются коллекцией 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 $y_1, ..., y_M$ – классы больных 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_m была отобрана выборка X_m, состоящая из случаев с подтвержденными эталонными паттернами данного заболевания. В свою очередь выборка X_0 представляет собой здоровых людей, не имеющих существенных отклонений от состояния нормы. В экспериментах для каждой болезни строилась двухклассовая выборка, содержащая больных X_m конкретным заболеванием и, одинаковых для всех, эталонных представителей здоровых людей X_0. Данную ценную информацию достаточно просто предоставить тематической модели. Предлагается создать дополнительный словарь W^c, содержащий всего 2 слова, характеризущих принадлженность человека к каждому из классов.

В таблице 1 приведены названия заболеваний, объемы выборок $|X|$ эталонных кардиограмм данного заболевания и объемы подвыборок $|X_m|$ кардиограмм больных без сопутствующих заболеваний.

Таблица 1: Двухклассовые обучающие выборки.

Болезнь		$\lvert X\rvert$	$\lvert X_m\rvert$
Абсолютно здоровые	АЗ	261	261
Миома матки	ММЭ	1022	761
Холецистит хронический	ХХЭ	784	523
Вегетососудистая дистония	ВДЭ	949	688
Язвенная болезнь	ЯБЭ	1027	766
Гипертоническая болезнь	ГБЭ	2211	1950
Ишемическая болезнь средца	ИБЭ	1564	1303
Сахарный диабет	СДЭ	1083	822
Мочекаменная болезнь	МКЭ	579	840
Желчнокаменная болезнь	ЖКЭ	644	383
Узловой зоб щитовиднойжелезы	УЩЭ	991	730
Хронический гастрит	ХГЭ	992	731
Дискинезия желчевыводящих путей	ДЖЭ	975	714
Метастазы	МТЭ	626	365
Рак общий	РОЭ	709	448
Аденома простаты	ЭАП	596	335
Аднексит хронический	ЭАХ	595	334
Анемия	ЭА	301	562
Гастродуоденит	ЭГД	545	284
Остеопороз	ОПЭ	505	244

3 Модели классификации

3.1 Вероятностная тематическая модель двухклассовой классификации

Пусть D – конечное множество (коллекция) текстовых документов, T – конечное множество тем. Обозначим через W объединение непересекающихся множеств W^{gram3} (словарь триграмм) и W^c (словарь меток классов). Каждый документ $d \in D$ представляет собой последовательность слов $w_1, \ldots, w_{n_d}$ из W, где n_d – длина документа. Принимая «гипотезу мешка слов», будем считать, что последовательность слов не важна, и учитывать только число вхождений n_{dw} слова w в документ d.

Вероятностная тематическая модель описывает условную вероятность появления метки класса c в документе $d \in D$ как вероятностную смесь распределений $\phi_{ct} = p(c|t)$ меток классов в темах $t \in T$ с коэффициентами $\theta_{td} = p(t|d)$, зависящими от документов:

$$p(c|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, \quad c \in W^c, \quad w \in W, \quad d \in D$$

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

В **вероятностном латентном семантическом анализе (PLSA)** [3] используется единственная модальность терминов (как правило, отдельных слов) и ставится задача максимизации логарифма правдоподобоия модели $p(c|d)$ при ограничениях неотрицательности и нормированности столбцов Φ и Θ:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \ \to \max_{\Phi, \Theta}; \tag{1}$$

$$\sum_{w \in W} \varphi_{wt} = 1; \qquad \varphi_{wt} \geq 0;$$

$$\sum_{t \in T} \theta_{td} = 1; \qquad \theta_{td} \geq 0.$$

В **мультимодальных моделях PLSA** [4] критерий логарифма правдоподобия вводится для каждой модальности и максимизируется их взвешенная сумма. В нашей задаче веса модальностей являются относительными, поэтому для триграмм возьмем

его равным 1.

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in W^{gram3}} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+\tau\sum_{d\in D}\sum_{w\in W^{c}} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}\ \to\max_{\Phi,\Theta};\qquad(2)$$

τ – вес модальности метки классов.

В **аддитивной регуляризации тематических моделей (ARTM)** к правдоподобию добавляются дополнительные критерии-регуляризаторы:

$$R(\Phi,\Theta)=\sum_{i=1}^{r}\tau_i R_i(\Phi,\Theta)\qquad L(\Phi,\Theta)+R(\Phi,\Theta)\ \to\max_{\Phi,\Theta};\qquad(3)$$

Полученная задача оптимизации решается с помощью регуляризованного EM-алгоритма [4, 5]. Вес модальности метки классов τ и коэффициенты регуляризации τ_i подбираются в эксперименте.

Регуляризатор разреживания ϕ_{wt}, $w\in W^{gram3}$ вводит в модель требование, чтобы каждая тема состояла из небольшого числа слов словаря триграмм, вероятности остальных слов в распределении ϕ_{wt} равны 0.

$$R(\Phi)=-\beta\sum_{t\in T}\sum_{w\in W^{gram3}}\beta_w\ln\phi_{wt}\ \to\max_{\Phi,\Theta};$$

Чем больше коэффициент регуляризации β, тем больше вероятностей ϕ_{wt} обращается в 0 на каждой итерации.

Высокая разреженность ϕ_{wt}, $w\in W^{gram3}$ повышает интерпретируемость модели, а также снижает размерность признакового пространства, осуществляя отбор признаков.

Регуляризатор декоррелирования ϕ_{wt}, $w\in W^{gram3}$ минимизирует ковариации между вектор-столбцами матрицы ϕ_{wt}, повышая различность тем и улучшая интерпретируемость модели.

$$R(\Phi)=-\alpha\sum_{t\in T}\sum_{s\in T\setminus t}\sum_{w\in W^{gram3}}\phi_{wt}\phi_{ws}\ \to\max_{\Phi,\Theta};$$

Декоррелирование приводит также к разреживанию тем и к более четкому выделению ядер тем, состоящих из слов w с сильно доминирующей вероятностью $p(t|w)$.

На практике два описанных регуляризатора показывают хорошие результаты, работая совместно [6].

3.2 Линейные модели классификации

Линейными моделями классификации называются диагностические правила, в которых решение принимается по значению взвешенной суммы признаков [7].

3.2.1 Наивный байесовский классификатор

Данная модель является упрощением тематической модели классификации.
Примем следующее наивное предположение: распределение меток классов внутри документа зависит только от распределения триграмм внутри данного документа и не зависит от самого документа.
Тогда, дискриминантная функция данной модели будет выглядеть следующим образом:

$$p(c|d) = \sum_{w \in W^{gram3}} p(c|w) p_{dw}$$

$$p_{dw} = \frac{n_{dw}}{n_d}$$

$$p(c|w) = \frac{n_{cw}}{n_w}$$

n_{dw} – частота триграммы w в документе d

n_d – длина документа d

n_{cw} – частота триграммы w в классе c

n_w – частота триграммы w во всей коллекции документов.

3.2.2 Синдромный алгоритм

Альтернативный вариант наивного байесовского классификатора основан на предположении, что диагностическую ценность имеют не частоты триграмм в документе, а только то, какие триграммы встречаются чаще, чем они могли бы встречаться случайно.

Согласно [2], каждое заболевание характеризуется своим набором информативных триграмм. Именно эти триграммы должны получать наибольшие по модулю веса в линейном классификаторе. Использование остальные «шумовых» триграмм может приводить к снижению информативности дискриминантной функции. Чтобы этого не происходило, предлагается ранжировать триграммы по убыванию некоторо-

го критерий информативности, и учитывать первые K триграмм. Число K является параметром метода и подбирается в эксперименте.

В работе [1] было рассмотрено несколько различных сочетаний типа используемых признаков, критерия отбора и формулы весов. Каждое сочетание формирует свою модель синдромного алгоритма. Мы же рассмотрим модель, показавшую лучшие результаты в ходе проведенных экспериментов:
Дискриминантная функция:

$$p(c|d) = \sum_{w \in W^{gram3}} \gamma_{cw}[p_{dw} \geq \theta]$$

- критерий отбора K признаков с наибольшими значениями:
  $B_w(X_m, \theta) = \frac{1}{|X_m|+2}(\sum_{d \in X_m}[p_{dw} \geq \theta] + 1)$
- формула весов признаков: $\gamma_{cw} = \ln \frac{B_w(X_m,\theta)(1-B_w(X,\theta))}{B_w(X,\theta)(1-B_w(X_m,\theta)}$
- θ подбирается в эксперименте.

4 Эксперименты

В данном разделе приводятся результаты обучения тематической модели, наивного байеса и синдромного алгоритма. Для настройки параметров и выбора лучшей версии каждого из алгоритмов используется 10×10-кратная кросс-валидация [7]. Итоговое качество диагностики оценивается по отложенной выборке.

4.1 Меры качества диагностики

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

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

Однако, данная мера не позволяет оценивать сами значения вероятностей, которые является очень важной характеристикой модели. Для этого введем дополнительную меру качества диагностики – меру LogLoss (Logistic Loss), позволяющей на качественном уровне оценивать значения вероятностей:

$$LogLoss = -[c = +1] \log p(c|d) - [c = -1] \log(1 - p(c|d))$$

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

- $\mathbf{p(w|t), w \in W^{gram3}}$.

  Эксперименты показывают возможность разреживания до 80 – 90% без потери конкурентноспособности модели. Таким образом, тематическая модель позволяет осуществлять отбор признаков.

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

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

  $$p(t) = \sum_{d \in X} p(t|d)p(d) \quad p(d) = \frac{1}{|X|} \quad p(c = +1) = \frac{|X_0|}{|X|}$$

  Согласно [2], каждое заболевание представляет собой набор *диагностических эталонов*. В настоящей работе *диагностическим эталоном* будем называть множество тем, определившихся в классе y_m. Эксперименты показывают однозначное выделение *диагностических эталонов* для каждой рассматриваемой болезни.

4.2 Тематическая модель

Данный раздел будет посвящен последовательным шагам становления конкурентноспособной и интерпретируемой тематической модели. Для этого рассмотрим ряд исследований, в котором каждый следующий эксперимент добавляет новые эвристики, позволяющие улучшать качество и интерпретируемость модели. Эксперименты проводились при помощи библиотеки BigARTM. Эталонная болезнь «Холецистит хронический» (ХХЭ) рассматривается в качестве демонстрационной. Поведение каждой

модели от итерации будем демонстрировать на графиках. Для мер качества AUC и LogLoss дополнительно будем показывать доверительный интервал 95%.

4.2.1 PLSA

Число тем: 2

Способ инициализации матрицы Φ: случайное распределение

Способ инициализации матрицы Θ: равномерное распределение

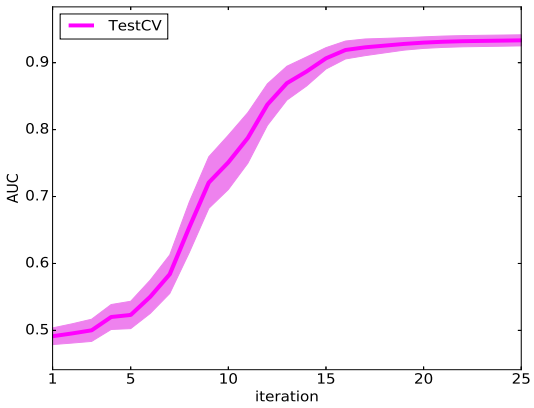

Рис. 1: AUC

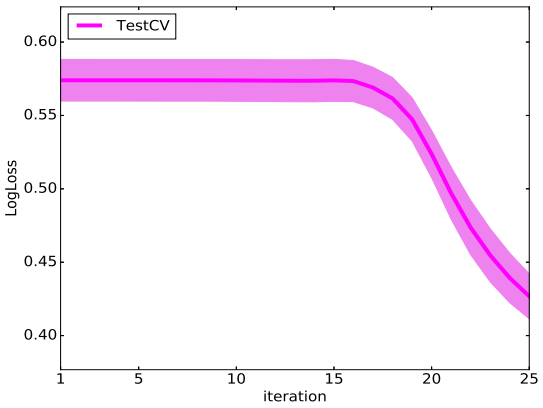

Рис. 2: LogLoss

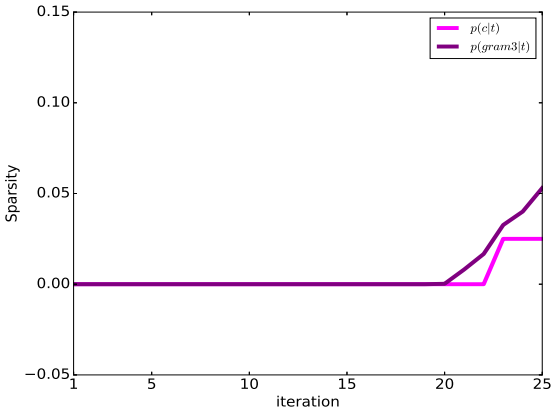

Рис. 3: Разреженность

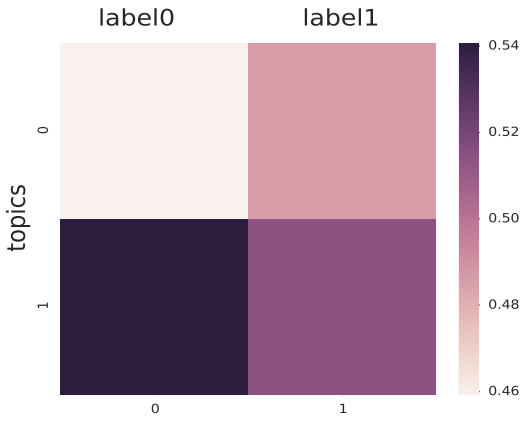

Рис. 4: $p(t|c)$

Тематическая модель по своей природе не является классификатором и не оп-

тимизирует в явном виде какую-либо метрику качества. Однако, она осуществляет мягкую кластеризацию, что позволяет нам использовать ее в задаче двухклассовой классификации.

Графики 1, 2 демонстрируют хорошее качество на последней итерации, однако, на начальных итерациях модель ведет себя, как случайный классификатор, что является следствием плохой инициализации. Также, наблюдаем почти полное отсутствие разреженности, что демонстрируют графики 3 и 4. Эталоны не выделились, отбор признаков не был осуществлен.

Итог:

+ Быстрая сходимость
+ Высокое качество на последних итерациях
– Низкое качество на первых итерациях
– Не интерпретируема

4.2.2 Подбор весов модальностей

Число тем: 2

Способ инициализации матрицы Φ: случайное распределение

Способ инициализации матрицы Θ: равномерное распределение

Вес модальности «метки классов» (τ): 10^3

Вес модальности «триграммы»: 1.0

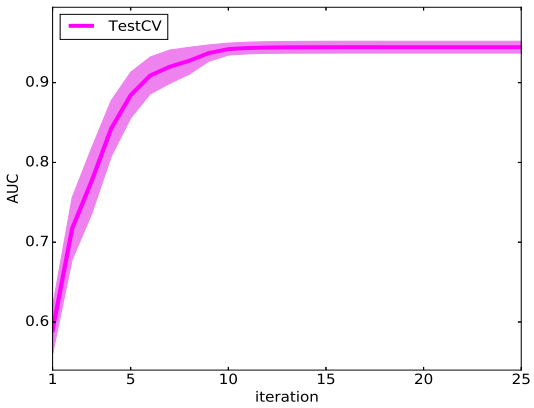

Рис. 5: AUC

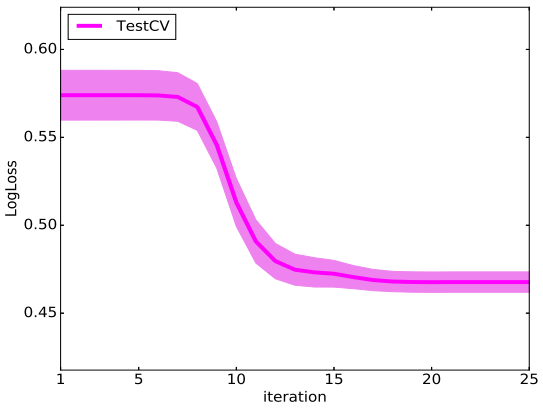

Рис. 6: LogLoss

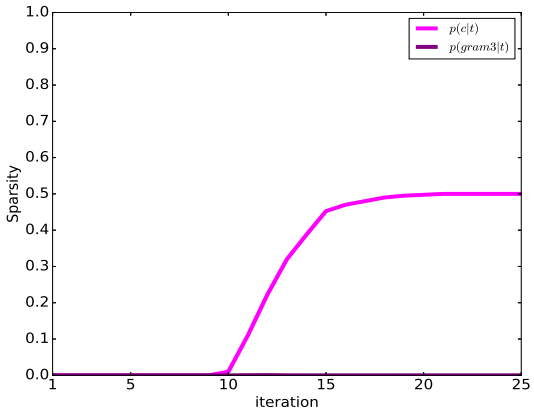

Рис. 7: Разреженность

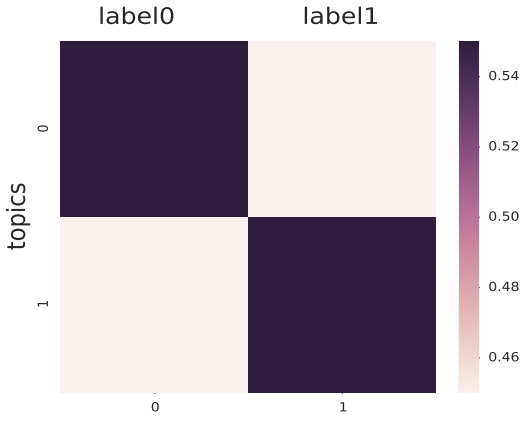

Рис. 8: $p(t|c)$

После определения модальностей, правдоподобие (2) разделилось на два слагаемых. Выбрав вес каждой части, мы указали тематической модели, какую модаль-

ность следует лучше описывать.

Оптимальное значение параметра τ выбиралось из заданной сетки значений в соответствии с критериями качества AUC и LogLoss. Для каждого значения веса проводилось 15 итераций EM-алгоритма. На каждой итерации производилось небольшое увеличение зафиксированного значения с целью избежания проседания качества на последующих итерациях. Из всех значений выбиралось то, при котором улучшался хотя бы один из критериев без существенного ухудшения другого [8].

В результате добились нужной степени разреженности матрицы $p(t|c)$ (графики 7, 8) без потери качества диагоностики (графики 5, 6).

Однако, данная эвристика не помогла осуществить отбор признаков, также как и существенно улучшить качество на начальных итерациях, хотя оно и стало выше по сравнению с предыдущей моделью. Это факт легко объяснить, интерпретируя второе слагаемое в правдоподобии (2) как наши априорные знания о решаемой задаче. Внесение данной информации в модель уже на начальных итерациях позволяет подняться выше случайного угадывания.

Итог:

+ Определились диагоностические эталоны
+ Быстрая сходимость
+ Высокое качество на последних итерациях
– Низкое качество на первых итерациях
– Отбор признаков не осуществлен

4.2.3 Подбор инициализации

В данном эксперименте предлагается опробовать различные эвристические способы инициализации матрицы Φ тематической модели.

Число тем: 8

Способ инициализации матрицы Θ: равномерное распределение

Вес модальности «метки классов» (τ): 10^3

Вес модальности «триграммы»: 1.0

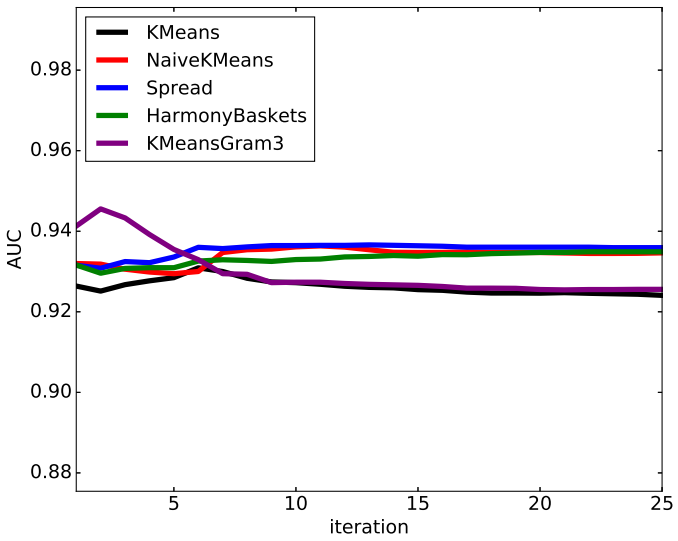

Рис. 9: AUC. Инициализация.

Так как тематическая модель осуществляет кластеризацию, логичнее всего ее инициализировать схожими по природе моделями.

Модель **KMeans**, кластеризующая документы, фактически реализует EM-алгоритм: на E-шаге осуществляем пересчет центров кластеров, на M-шаге пересчитываем кластеры по степени близости к центру.

Наивный KMeans (**NaiveKMeans**), в отличие от обычного, пытается найти два самых удаленных объекта ото всех остальных. Принимая наивное предположение, что таким образом мы сможем найти объекты, лучше всего описывающие распределение всей выборки.

Метод **Spread**, в литературе «Растопырка», является более продвинутой вер-

сией **NaiveKMeans**. В нем мы находим объекты, максимально «растопырившиеся» относительно всей выборки, и называем их центрами.

Метод **HarmonyBaskets** основан на вычислении коэффициента гармонии (КГ), являющего собой отношение числа A, B, E, F к числу C, D в кодограмме. Кластер образуют объекты с близким значением КГ.

Метод **KMeansGram3** кластеризует по триграммам. Отсутствующие триграммы в каждом классе заполняются равномерным распределением.

Самое высокое качество на начальной итерации показывает **KMeansGram3**, однако, вместе с обычным **KMeans**, начинает проседать от итерации (график 9).

Как можно видеть **Spread**, **NaiveKMeans** и **HarmonyBaskets** хорошо вливаются в работу тематической модели.
Лучшим является метод **Spread**. Оценим итоговую модель с выбранной нами инициализацией.

Как можно видеть, начальная инициализация сумела оказать влияние не только на значение AUC (график 10), но позволила осуществить небольшой отбор признаков, оставив порядка 80% информативных триграмм (график 12). Это, конечно, не является требуемый результатом, однако представляет собой интересный факт поведения модели.

График 13 показывает, что выбранное нами интуитивное число тем не очень соответствует данной болезни. Несколько тем $(4, 5, 6)$ кажутся лишними, возможно описывающими другие болезни настоящей выборки. Выбором оптимального числа тем займемся в следующих экспериментах.

Итог:

\+ Хорошее качество на первых итерациях в смысле меры AUC

\+ Определились диагоностические эталоны

\+ Быстрая сходимость

\+ Высокое качество на последних итерациях

\+− Частичный отбор признаков

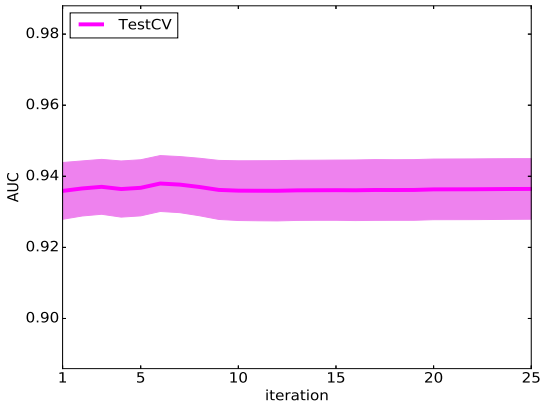

Рис. 10: AUC

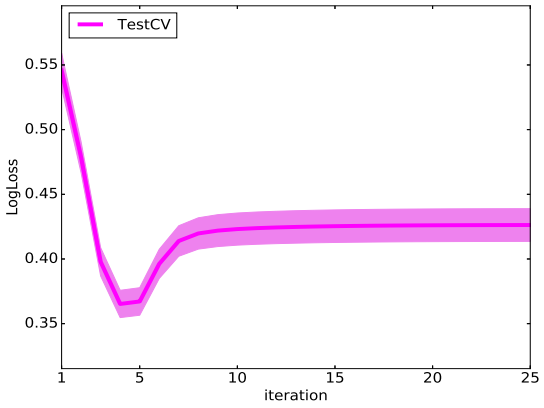

Рис. 11: LogLoss

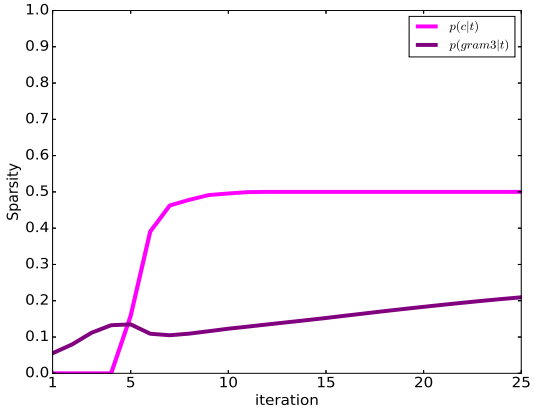

Рис. 12: Разреженность

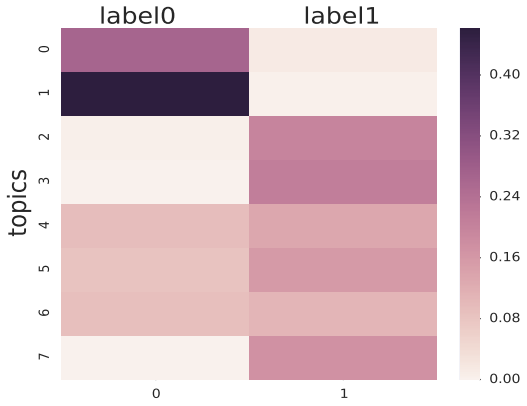

Рис. 13: $p(t|c)$

4.2.4 Подбор регуляризации

Число тем: 8

Способ инициализации матрицы Φ: **Spread**

Способ инициализации матрицы Θ: равномерное распределение

Вес модальности «метки классов» ($\boldsymbol{\tau}$): 10^3

Вес модальности «триграммы»: 1.0

Сила декорреляции распределения триграмм в темах ($\boldsymbol{\alpha}$): 1.0

Сила разреживания распределения триграмм в темах ($\boldsymbol{\beta}$): -10.0

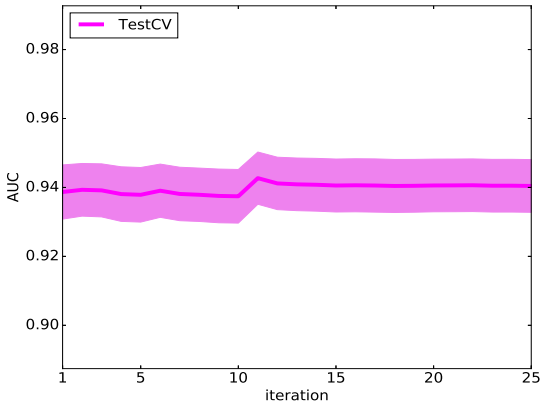

Рис. 14: AUC

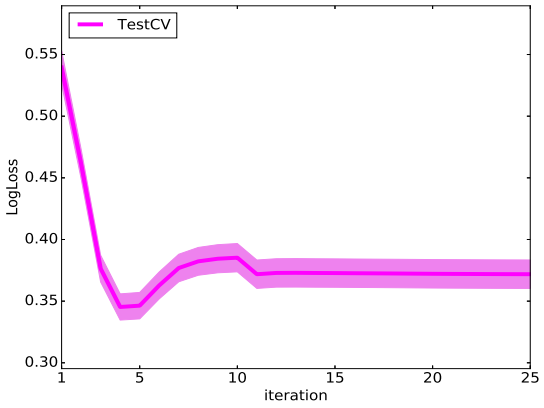

Рис. 15: LogLoss

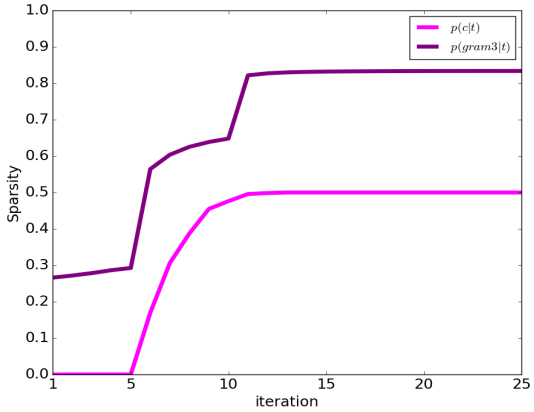

Рис. 16: Разреженность

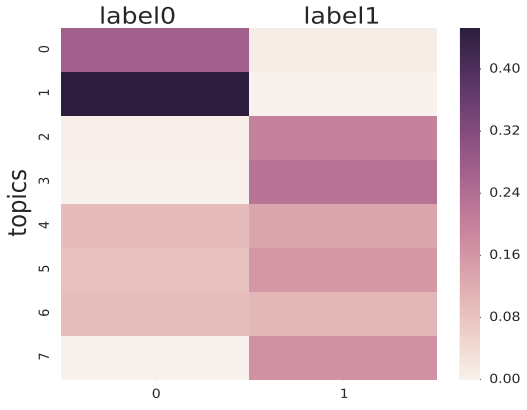

Рис. 17: $p(t|c)$

В данном эксперименте предлагается включить два регуляризатора: декорреляцию и разреживание распределения $p(w|t)$, $w \in W^{gram3}$. 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. Для каждого значения коэффициента регуляризации производилось 15 итераций EM-алгоритма. Первые 5 итераций включалась только декорреляция, затем 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.

Декоррелирующий регуляризатор обладает побочным эффектом разреживания, поэтому уже на первой итерации был произведен отбор признаков, оставивший порядка 70% информативных триграмм. Добавление разреживающего регуляризатора, начиная с 10 итерации, осуществило требуемый отбор признаков, оставив лишь 20% информативных триграмм (график 16). Качество модели продолжает быть высоким (графики 14, 15).

Итог:

+ Отбор признаков
+ Определились диагностические эталоны
+ Быстрая сходимость
+ Высокое качество на последних итерациях
+ Хорошее качество на первых итерациях в смысле меры AUC

4.2.5 Умная стратегия подбора параметров

Число тем: 4

Способ инициализации матрицы Φ:

$0.33 * \textbf{KMeans} + 0.33 * \textbf{Spread} + 0.33 * \textbf{HarmonyBaskets}$

Способ инициализации матрицы Θ: равномерное распределение

Вес модальности «метки классов» ($\boldsymbol{\tau}$): 10^4

Вес модальности «триграммы»: 1.0

Сила декорреляции распределения триграмм в темах ($\boldsymbol{\alpha}$): 1.0

Сила разреживания распределения триграмм в темах ($\boldsymbol{\beta}$): -100.0

Алгоритм умной стратегии подбора всех исследуемых параметров можно условно поделить на три последовательных этапа. Каждый следующий этап фиксирует лучшие параметры, найденные в ходе предыдущих экспериментов. Оптимальными назначаются те значения параметров, при которых улучшается хотя бы один из критериев качества AUC или LogLoss без существенного ухудшения другого.

Первый этап заключается в подборе оптимальных числа тем и комбинации инициализаций. Для каждого фиксированного значения числа тем последовательно перебираем каждую инициализацию, фиксируем лучшую, добавляем к ней в выпуклую комбинацию еще одну, фиксируем лучшую и так далее. Для каждого выбранного набора производилось 2 итерации EM-алгоритма.

Второй этап заключался в выборе оптимального значения параметра $\boldsymbol{\tau}$. Также накладывалось дополнительное требование – четкое выделение диагностических эталонов в каждом классе. Производилось 15 итераций для каждого фиксированного значения.

Третий этап обеспечивал подбор $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$. В качестве дополнительного требования выступало ограничение на количество используемых признаков модели – не больше 20%. Для каждого выбранного набора производилось 15 итерации EM-алгоритма.

В результате, качество в смысле AUC изменилось не сильно (график 18), зато в смысле LogLoss наблюдаем улучшение уже во втором знаке после запятой (график 19).

Произведен качественный отбор признаков, позволивший оставить всего лишь

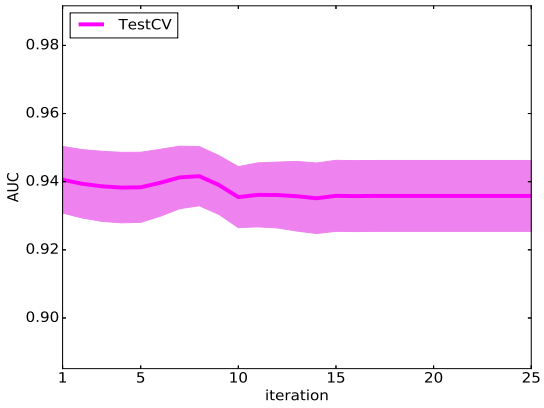

Рис. 18: AUC

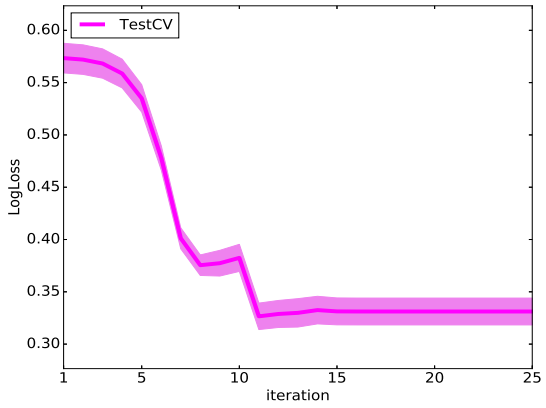

Рис. 19: LogLoss

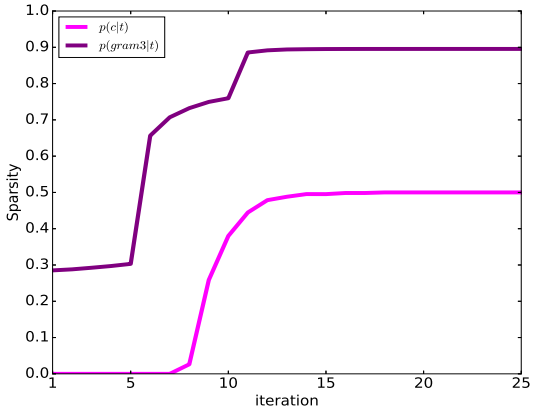

Рис. 20: Разреженность

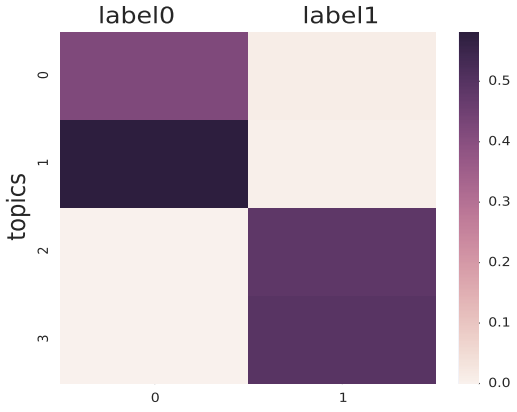

Рис. 21: $p(t|c)$

10% информативных триграмм без существенной потери качества диагностики (график 20). В каждом классе однозначным образом выделились диагностические эталоны (графики 20, 21).

В таблицах 2 и 3 приведены значения AUC и LogLoss на тестовых выборках всех представленных болезней для каждой из рассмотренных выше тематических моделей. Видно, что каждая последующая эвристика осуществляет более качественную диагностику хотя бы по одному из критериев.

Каждое достижение тематической модели выделено цветом, яркость которого говорит о значимости соответствующего достижения.

Модель **MultiPLSA** позволила однозначно выделить диагностические эталоны в каждом классе. Модель **ARTM** осуществила качественный отбор признаков.

И, наконец, модель **ARTM_smart** суммаризовала все предыдущие шаги к успеху в единый алгоритм умной стратегии подбора всех исследуемых параметров, позволяющий для абсолютно новых выборок болезней строить конкурентноспособные и интерпретируемые тематические модели.

	PLSA	MultiPLSA	MultiPLSA + smart_init	ARTM	ARTM_smart
ММЭ	0.8896	0.8907	0.8911	0.8739	0.8915
ХХЭ	0.8954	0.9041	0.9034	0.9007	0.8878
ВДЭ	0.7669	0.8324	0.8031	0.8122	0.8251
ГДЭ	0.8949	0.9188	0.912	0.906	0.9097
ЯБЭ	0.9039	0.9092	0.9111	0.8963	0.8951
ГБЭ	0.8555	0.9217	0.9157	0.9019	0.9077
ИБЭ	0.9238	0.94	0.9381	0.9251	0.9255
СДЭ	0.9303	0.9419	0.9386	0.9356	0.911
МКЭ	0.8982	0.9057	0.8993	0.8937	0.897
ЖКЭ	0.9549	0.962	0.9602	0.9599	0.957
УЩЭ	0.9052	0.924	0.9223	0.9143	0.9194
ХГЭ	0.9208	0.9332	0.9296	0.9241	0.9235
ДЖЭ	0.8994	0.9068	0.9031	0.8958	0.8835
РОЭ	0.9183	0.935	0.9343	0.9318	0.9181
ЭАП	0.9372	0.9547	0.9504	0.9498	0.9534
ЭАХ	0.8311	0.8585	0.855	0.8578	0.854
ЭА	0.8336	0.8828	0.8735	0.8732	0.8739
ГБК	0.9316	0.9526	0.9636	0.9478	0.9458
ТБЦ	0.9073	0.9338	0.9247	0.9147	0.9136
ОПЭ	0.9738	0.9794	0.9829	0.9778	0.9754

Таблица 2: Значения AUC на тестовых выборках при использовании разных тематических моделей

	PLSA	MultiPLSA	MultiPLSA + smart_init	ARTM	ARTM_smart
ММЭ	0.434	0.5297	0.4721	0.4176	0.4002
ХХЭ	0.4743	0.5139	0.4913	0.4221	0.4176
ВДЭ	0.5565	0.5898	0.5501	0.5061	0.4849
ГДЭ	0.4671	0.518	0.557	0.4494	0.4279
ЯБЭ	0.4215	0.513	0.481	0.4384	0.4249
ГБЭ	0.4359	0.466	0.3316	0.2795	0.288
ИБЭ	0.5352	0.4436	0.3581	0.3146	0.314
СДЭ	0.5949	0.4615	0.4183	0.3633	0.3865
МКЭ	0.442	0.5148	0.5065	0.4314	0.4283
ЖКЭ	0.4059	0.4583	0.5191	0.358	0.3487
УЩЭ	0.6243	0.4933	0.4569	0.3978	0.4642
ХГЭ	0.6201	0.4913	0.4582	0.3846	0.3795
ДЖЭ	0.4801	0.5163	0.4895	0.4331	0.411
РОЭ	0.4121	0.4762	0.5175	0.3738	0.4121
ЭАП	0.4197	0.4712	0.5555	0.3807	0.3851
ЭАХ	0.5591	0.5643	0.5999	0.5101	0.5031
ЭА	0.6154	0.553	0.6037	0.4788	0.4473
ГБК	0.4633	0.4813	0.5135	0.3883	0.381
ТБЦ	0.5075	0.5034	0.5424	0.4258	0.4585
ОПЭ	0.3653	0.3967	0.462	0.302	0.3065

Таблица 3: Значения LogLoss на тестовых выборках при использовании разных тематических моделей

4.3 Синдромный алгоритм

В данном разделе предлагается оценить работу синдромного алгоритма в зависимости от числа используемых признаков. Эталонная болезнь «Холецистит хронический» (ХХЭ) взята в качестве демонстрационной.

Так как метод по своей природе не является вероятностным будем калибровать его выход при помощи калибровки-Платта [9].

На графиках 22, 23 показаны значения AUC и LogLoss соответственно на обучающих и тестовых выборках кросс-валидации в зависимости от значений параметра K.

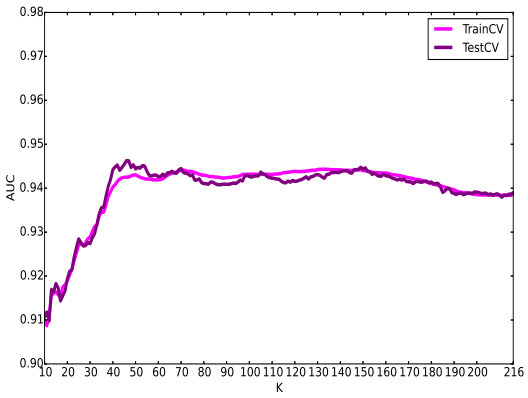

Рис. 22: AUC

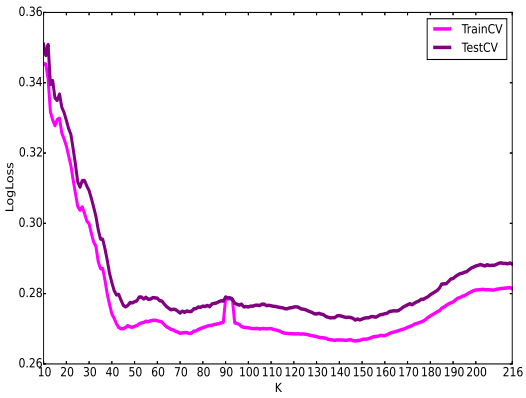

Рис. 23: LogLoss

В целом, синдромный алгоритм не переобучается, однако, использование более 70% признаков приводит к ухудшению качества. Модель становится более сложной, возможно влияние шумовых признаков.

4.4 Сравнение методов

В таблицах 4, 5 приведены значения AUC и LogLoss соответственно на тестовых выборках всех представленных болезней для следующих моделей: **ARTM_smart**, **NB** – наивный байесовский классификатор и **SA** – синдромный алгоритм с подобранным значением параметра K.

	NB	ARTM_smart	SA
ММЭ	0.8837	0.8915	0.916
ХХЭ	0.8965	0.8878	0.9201
ВДЭ	0.8233	0.8251	0.8261
ГДЭ	0.9144	0.9097	0.9602
ЯБЭ	0.9038	0.8951	0.9465
ГБЭ	0.9085	0.9077	0.9531
ИБЭ	0.9289	0.9255	0.97
СДЭ	0.937	0.911	0.9434
МКЭ	0.903	0.897	0.9176
ЖКЭ	0.9612	0.957	0.9764
УЩЭ	0.9165	0.9194	0.9322
ХГЭ	0.9242	0.9235	0.9309
ДЖЭ	0.8985	0.8835	0.9217
РОЭ	0.9296	0.9181	0.941
ЭАП	0.9521	0.9534	0.9475
ЭАХ	0.8641	0.854	0.8566
ЭА	0.8837	0.8739	0.8875
ГБК	0.9535	0.9458	0.9747
ТБЦ	0.9275	0.9136	0.9769
ОПЭ	0.9773	0.9754	0.9827

Таблица 4: AUC

	NB	ARTM_smart	SA
ММЭ	0.3811	0.4002	0.3422
ХХЭ	0.3931	0.4176	0.3455
ВДЭ	0.4893	0.4849	0.4804
ГДЭ	0.3623	0.4279	0.2498
ЯБЭ	0.3714	0.4249	0.3034
ГБЭ	0.2499	0.288	0.1972
ИБЭ	0.2722	0.314	0.188
СДЭ	0.3158	0.3865	0.2834
МКЭ	0.3672	0.4283	0.3485
ЖКЭ	0.2768	0.3487	0.2082
УЩЭ	0.3576	0.4642	0.3252
ХГЭ	0.3236	0.3795	0.32
ДЖЭ	0.3719	0.411	0.3274
РОЭ	0.3321	0.4121	0.3037
ЭАП	0.313	0.3851	0.2922
ЭАХ	0.4686	0.5031	0.4712
ЭА	0.4475	0.4473	0.4496
ГБК	0.2844	0.381	0.1948
ТБЦ	0.3313	0.4585	0.1904
ОПЭ	0.207	0.3065	0.1769

Таблица 5: LogLoss

Модель **ARTM_smart** учитывала особенности каждой болезни, подбирая оптимальное число тем для каждого класса. В большинстве своем более высокое качество классификации демонстрировали модели с 2, 4 или 8 темами. То есть тематическая модель по сложности была близка основному baseline-алгоритму настоящей работы – модели **NB**. Накладывая дополнительные условия разреженности, тематическая модель была подвержена еще более сложному испытанию – отбору признаков и выделению диагностических эталонов, позволяющих выявлять паттерны конкретных заболеваний.

Данным свойством не обладает ни один из конкурирующих методов. Качество диагностики и вычислительная сложность всех рассмотренных моделей являются сопоставимыми.

5 Заключение

Данная работа иллюстрирует применение нового подхода в информационном анализе электрокардиосигналов – использование регуляризованной тематической модели классификации. Пройден путь становления интерпретируемой и конкурентноспособной тематической модели, позволяющей осуществлять отбор признаков, а также поиск кодовых образов заболеваний. Предложен алгоритм, обеспечивающий для любого заболевания автоматический подбор параметров тематической модели, оптимизирующий совокупность критериев качества.

Высокое качество классификации при обязательном требовании интерпретируемости позволяет использовать тематическую модель в качестве качественной диагностической системы различных заболеваний.

6 Список литературы

[1] Воронцов К.В., Целых В.Р. 2015. Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов.

[2] Успенский В.М. 2008. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализ электрокардиосигналов.

[3] Hofmann, T. 1999. Probabilistic latent semantic indexing. Proceedings of the 22nd annual in- ternational ACM SIGIR conference on Research and development in information retrieval.

[4] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova et al. 2015. Non-bayesian addi- tive regularization for multimodal topic modeling of large collections. Proceedings

of the 2015 Workshop on Topic Models: Post-Processing and Applications.

[5] Vorontsov, K. V., and A. A. Potapenko. 2015. Additive regularization of topic models. Ma- chine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications.

[6] Vorontsov, K. V., and A. A. Potapenko. 2014. Tutorial on probabilistic topic modeling: Ad- ditive regularization for stochastic matrix factorization. AIST'2014, Analysis of Images, Social networks and Texts.

[7] Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин).

[8] Янина А.О., Воронцов К.В. 2016. Мультимодальные тематические модели для-разведочного поиска в коллективном блоге.

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