Московский государственный университет
имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Шапулин Андрей Валентинович

Регуляризация вероятностных тематических моделей для классификации символьных последовательностей

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

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

Москва, 2015

Содержание

1 Введение

Тематическое моделирование является инструментом статистического анализа текстов и предназначено для выявления латентной тематики коллекций текстовых документов. Тематическая модель представляет каждую тему в виде дискретного распределения на множестве слов, а каждый документ — в виде дискретного распределения на множестве тем.

Задача поиска такого представления имеет бесконечно много решений и является некорректно поставленной. Для повышения устойчивости решения к модели предъявляются дополнительные требования. Аддитивная регуляризация тематических моделей (APTM) позволяет комбинировать требования в произвольных сочетаниях путём оптимизации логарифма правдоподобия модели с линейной комбинацией регуляризаторов [3].

В последнее время тематическое моделирование всё чаще применяется при анализе сигналов, изображений и видеопоследовательностей — в областях, далёких от обработки естественного языка. В данной работе описываются методы применения тематического моделирования для классификации символьных последовательностей.

2 Основная часть

2.1 Векторизация символьной последовательности

Для того, чтобы применить к некоторой символьной последовательности инструментарий тематического моделирования, необходимо провести процесс векторизации.

Рассмотрим символьную последовательность $S = \{s_k\}_{k=1}^{N}$ букв из алфавита $\mathcal{A}$. Тогда назовем *n-граммой* слово, которое образовано n последовательными буквами S: $s_p, \ldots, s_{p+n-1}$. Пусть в алфавите $|\mathcal{A}|$ различных букв, тогда множество всех различных n-грамм W имеет мощность $|W| = |\mathcal{A}|^n$. Далее определим частоту n-граммы p_w, как отношение числа ее вхождений n_w в последовательность S к общему числу n-грамм, которое равно $N - n + 1$:

$$n_w = \sum_{k=1}^{N-n+1} \prod_{i=0}^{n-1} [s_{k+i} = w_i], \quad p_w = \frac{n_w}{N - n + 1}.$$

На рис. 1, 2 показан пример формирования векторного представления символьной последовательности.

Рис. 1: Пример формирования триграмм.

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

Рис. 2: Векторное представление символьной последовательности.

2.2 Теория тематических моделей

Кратко опишем задачу тематического моделирования с классификацией. Подробное описание теории можно найти в статье К. В. Воронцова [3].

Пусть мы имеем коллекцию документов D, множество слов в коллекции W и множество тем T. Нам необходимо получить распределения $\varphi_{wt} = p(w \mid t)$ для всех тем $t \in T$ и распределения $\theta_{td} = p(t \mid d)$ для всех документов $d \in D$. При этом используется основная гипотеза условной независимости $p(w \mid d, t) = p(w \mid t)$. Далее по формуле полной вероятности:

$$p(w \mid d) = \sum_{t \in T} p(t \mid d) p(w \mid t)$$

Формулы для EM-алгоритма выводятся из максимизации логарифма

правдоподобия

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

с ограничениями неотрицательности и нормировки

$$\varphi_{wt} > 0; \ \sum_{w \in W} \varphi_{wt} = 1; \ \theta_{td} > 0; \ \sum_{t \in T} \theta_{td} = 1$$

где n_{dw} — число вхождений термина w в документ d. Задача решается при помощи EM-алгоритма. На E-шаге вычисляются $H_{dwt} = p(t \mid d, w)$. Формулы M-шага в данном случае имеют вид:

$$\varphi_{wt} \propto n_{wt} \qquad \theta_{td} \propto n_{dt}$$

2.2.1 Регуляризация тематических моделей.

Далее в [3] было предложено ввести регуляризацию тематической модели. Для этого необходимо ввести в модель регуляризаторы R_i и максимизировать их линейную комбинацию с логарифмом правдоподобия L:

$$R(\Phi, \Theta) = \sum_{i}^{n} \tau_i R_i$$

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

с ограничениями неотрицательности и нормировки столбцов Θ и Φ. С введением регуляризаторов меняются формулы M шага:

$$\varphi_{wt} \propto (n_{wt} + \varphi_{wt} \frac{\partial R(\Phi, \Theta)}{\partial \varphi_{wt}})_+ \qquad \theta_{td} \propto (n_{dt} + \theta_{td} \frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}})_+$$

2.2.2 Модель классификации.

Далее рассмотрим задачу классификации документов. Пусть каждому документу соответствует набор некоторых классов $c \in C$. В [3] предлагаются различные модели классификации документов: моделирование классов темами, моделирование классов распределениями тем и тематическая модель классификации. В этой работе рассматривается последняя модель. Будем выражать распределение классов документов $p(c \mid d)$ через распределение тем документов $\theta_{td} = p(t \mid d)$.

$$p(c \mid d) = \sum_{t \in T} p(c \mid t) p(t \mid d) = \sum_{t \in T} (\psi_{ct} \theta_{td})$$

Это возможно благодаря использованию гипотезы условной независимости

$$p(c \mid t, d) = p(c \mid t),$$

означающей, что для классификации документа d достаточно знать только его тематику. Модель обучается по заданной матрице m_{dc}, которая описывает информацию о принадлежности документа d некоторому классу c. Она может быть как логической (принадлежит или нет) так и частотной (на сколько документ d принадлежит конкретному классу c). Задача решается также при помощи EM-алгоритма. На выходе мы имеем матрицы φ_{wt}, θ_{td}, ψ_{ct}. Далее для классификации новой (тестовой) коллекции мы запускаем EM-алгоритм без классификации, получаем матрицу θ_{td}^{test} и считаем искомую вероятность отнесения документа d к классу c, как $p(c \mid d) = \sum_t \psi_{ct}\theta_{td}^{test}$. Для этой тематической модели в [3] предложено использовать регуляризатор классификации:

$$R(\Psi, \Theta) = \sum_{d \in D} \sum_{c \in C} m_{dc} \ln(\sum_{t \in T} \psi_{ct}\theta td)$$

На E-шаге дополнительно вычисляется вероятность $H'_{dct} = p(t \mid d, c)$. Формулы для M-шага принимают вид:

$$\theta_{td} \propto n_{dt} + \tau m_{dt} \qquad \psi_{ct} \propto m_{ct}$$

где $m_{dt} = \sum_{c \in C} m_{dc} H'_{dct} \qquad m_{ct} = \sum_{d \in D} m_{dc} H'_{dct}$

2.2.3 Разреживающе-сглаживающие регуляризаторы.

Гипотеза разреженности говорит о том, что каждый документ d и каждый термин w связан с небольшим числом тем t. Поэтому большая часть вероятностей $p(t \mid d)$ и $p(w \mid t)$ должна обращаться в нуль. Сглаживающий регуляризатор также называется регуляризатором Дирихле, т. к. берет свое начало из модели Latent Dirichlet Allocation (LDA). [3] Вывод регуляризаторов Φ и Θ приводится в [3], формулы для M-шага модифицируются следующим образом:
Для разреживающего резуляризатора

$$\theta_{td} \propto (n_{dt} - \alpha_t)_+ \qquad \varphi_{wt} \propto (n_{wt} - \beta_w)_+$$

Для сглаживающего регуляризатора

$$\theta_{td} \propto (n_{dt} + \alpha_t)_+ \qquad \varphi_{wt} \propto (n_{wt} + \beta_w)_+$$

где $\alpha_t > 0$; $\beta_w > 0$. Основываясь на этих регуляризаторах можно построить их комбинацию для того чтобы некоторые темы сглаживать, а некоторые разреживать. Например, у нас есть фоновые темы $t \in B$, которые надо сгладить, а некоторые темы разрежены $t \in S$. Полученные формулы имеют вид:

$$\theta_{td} \propto (n_{dt} + [t \in B]\alpha_t - [t \in S]\alpha_t)_+ \qquad \varphi_{wt} \propto (n_{wt} + [t \in B]\beta_w - [t \in S]\beta_w)_+$$

2.2.4 Частотный регуляризатор.

Для классификации с несбалансированными классами хорошо подходит частотный регуляризатор. Будем минимизировать дивергенцию Кульбака-Лейблера между оценкой безусловного распределения классов по тематической модели $p(c)$ и наблюдаемыми частотами классов $\hat{p}(c)$ Что эквивалентно максимизации

$$\sum_{c \in C} \hat{p}(c) \ln p(c) = \left\{ \; p(c) = \sum_{t \in T} p(c \mid t) p(t) \; \right\} = \sum_{c \in C} \frac{|D_c|}{|D|} \ln \psi_{ct} p(t)$$

$$R(\Psi) = \tau \sum_{c \in C} |D_c| \ln \psi_{ct} p(t) \to \max$$

где $p(t) = \frac{n_t}{n}$, $|D_c|$ – количество документов, принадлежащих классу c. Получаем формулы М-шага:

$$\psi_{ct} \propto m_{ct} + \tau \frac{|D_c| \psi_{ct} m_t}{\sum_{s \in T} \psi_{cs} m(s)}$$

3 Эксперимент

3.1 Описание и цели

Профессором Успенским [4] было установлено, что генерируемые сердцем сигналы могут описывать состояние здоровья человека. Эту информацию можно использовать для диагностики самых различных заболеваний, а не только заболеваний сердечно-сосудистой системы.

Целью данной работы являлось построение и исследование модели тематической классификации для решения задачи диагностики заболеваний с учетом ее специфики. Необходимо было применить существующую теорию тематического моделирования, опираясь на уже известные факты и различные гипотезы, характеризующие эту задачу.

Информационный анализ электрокардиосигналов основан на преобразовании электрокардиосигнала в так называемую кодограмму — символьную последовательность, кодирующую знаки приращений интервалов и амплитуд R-зубцов [4],[2]. В терминах тематических моделях кодограмму можно рассматривать как текстовый документ, а короткие подпоследовательности символов — как слова. Для каждой кодограммы имеется список установленных диагнозов обследуемого, и задача заключается в построении алгоритма диагностики. Для решения данной задачи в [2] использовались линейные методы классификации с отбором признаков. В терминах тематического моделирования это эквивалентно предположению, что каждое заболевание хорошо описывается одной темой — вероятностным распределением на множестве слов. Высокие уровни чувствительности и специфичности позволяют утверждать, что генерируемые сердцем сигналы несут существенную информацию для диагностики различных заболеваний внутренних органов, причём не только сердечно-сосудистой системы.

В данной работе ставится задача проверить более общее предположение, что каждое заболевание может быть ещё лучше описано вероятностной смесью нескольких тем. Тематические модели успешно применялись в [1] для решения сложных задач классификации с большим числом несбалансированных, пересекающихся и взаимозависимых классов. Однако непосредственное применение таких моделей к кодограммам ЭКГ не дало выигрыша качества классификации по сравнению с линейными моделями SVM, логистической регрессии и даже наивного байесовского классификатора.

Более успешным оказалось применение APTM с комбинацией регуляризаторов. Регуляризатор разреживания позволяет находить диагностические эталоны каждого заболевания — темы, состоящие из небольшого числа слов. Регуляризатор сглаживания позволяет выделять фоновую тему, не специфичную ни для какого заболевания (в обработке естественного языка аналогом является выделение общей лексики языка). Частотный регуляризатор помогает побороть проблему несбалансированных классов.

3.2 Исходные данные и условия эксперимента

Рассмотрим подробнее процесс преобразования кардиограмм в символьные последовательности. Сначала на кардиограмме выделяются так называемые *кардиоциклы*, которые являются периодами кардиосигнала. Утверждается [4], что основную информационную ценность электрокардиограммы несут в себе интервалы T_n и амплитуды R_n n-го кардиоцикла, а точнее их приращения $dR_n = R_{n+1} - R_n$ и $dT_n = T_{n+1} - T_n$. Также вводится аналог фазового угла $\alpha_n = \operatorname{arctg} R_n/T_n$. Пример размеченной электрокардиограммы изображен на рис. 3.

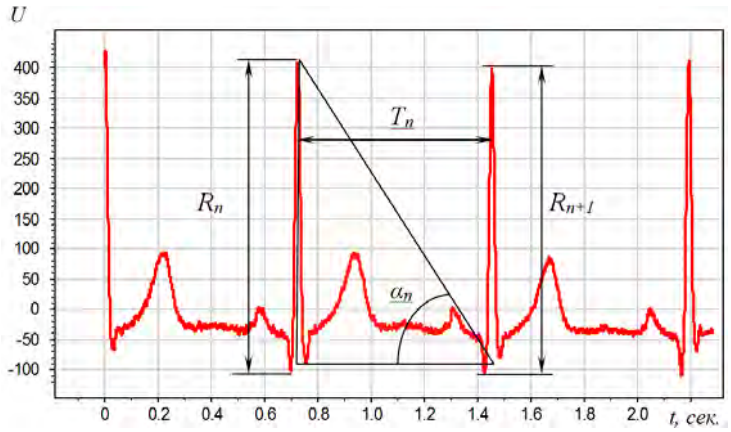


Рис. 3: Электрокардиограмма. Амплитуды R_n и интервалы T_n n-го кардиоцикла.

В электрокардиограмме возможно появление только 6 сочетаний знаков величин dT_n, dR_n, $d\alpha_n$ (таблица 1).

Таблица 1: Возможные знаки приращений.

	A	B	C	D	E	F
dR_n	+	-	+	-	+	-
dT_n	+	-	-	+	+	-
$d\alpha_n$	+	+	+	-	-	-

Поэтому их можно закодировать буквами алфавита из 6-ти символов $\mathcal{A} = \{A, B, C, D, E, F\}$. Данный процесс получения символьной последовательности (кодограммы) из непрерывного сигнала называется *дискретизацией*.

Эксперименты проводились на выборке 11 894 кодограмм с диагнозами по 18 заболеваниям. Данные собирались при помощи диагностической системы «Скринфакс» [4]. Каждому заболеванию соответствует выборка кодограмм, для которых при помощи клинического анализа был надежно установлен диагноз. Также была тщательно отобрана выборка абсолютно здоровых людей, у которых не было выявлено отклонений от нормы. Данные по заболеваниям приведены в таблице 2.

Таблица 2: Используемые заболевания. Аббревиатура и объем выборки.

анемия железодефицитная	ЖДА	260
аденома простаты	ДГПЖ	260
аднексит хронический	АХ	276
вегетососудистая дистония	ВСД	694
гипертоническая болезнь	ГБ	1894
асептический некроз головки бедренной кости	НГБК	324
хронический гастрит (гастродуоденит) гиперацидный	ХГ1	324
хронический гастрит (гастродуоденит) гипоацидный	ХГ2	700
дискинезия желчевыводящих путей	ДЖВП	717
желчнокаменная болезнь	ЖКБ	278
ишемическая болезнь сердца	ИБС	1265
мочекаменная болезнь	МКБ	654
миома матки	ММ	781
рак общий (онкопатология различной локализации)	РО	530
сахарный диабет (СД1 и СД2)	СД	871
узловой (диффузный) зоб щитовидный железы	УЩ	748
холецистит хронический	ХХ	340
язвенная болезнь	ЯБ	785
абсолютно здоровые	**АЗ**	**193**

3.3 Методы

Для каждой отдельно взятой болезни решается задача бинарной классификации: отделение класса абсолютно здоровых людей от класса людей, больных данной болезнью. Таким образом решается 18 задач классификации на два класса.

Частотные матрицы генерируются из кодограмм больных соответствующей болезнью при помощи рассмотренного выше алгоритма векторизации с использованием n-грамм длины 3 (триграмм).

Итого на вход алгоритма подается обучающая частотная матрица n_{dw}^k, где документы d – кодограммы, слова w - триграммы. Также для каждой кодограммы d, соответствующей некоторому пациенту, есть размеченная врачом информация (диагноз) болен человек – 1 или здоров – 0.

Анализ качества модели производится по 10-блочной кросс-валидации. Выборка случайным образом делилась на 10 равных частей и каждую итерацию одна часть использовалась как тестовая, а объединение остальных, как обучающая выборка. По тестовой выборке вычисляются следущие оценки качества модели:

- Чувствительность (TPR)
- Специфичность (TNR)
- AUC

Обозначим результат работы классификатора болезни k на новом объекте d, как $a_k(d)$.

Чувствительность определяется, как доля больных, предсказанных классификатором, среди истинно больных.

$$\text{чувствительность} = \frac{1}{|X_k|} \sum_{d \in X_k} [a_k(d) = 1],$$

где X_k — множество людей, действительно больных болезнью k. Специфичность определяется как доля предсказанных здоровых, среди истинно здоровых.

$$\text{специфичность} = \frac{1}{|X_0|} \sum_{d \in X_0} [a_k(d) = 0],$$

где X_0 — множество абсолютно здоровых людей.

Так как алгоритм классификации возвращает распределение классов в документах, то варьируя порог отнесения к классу больных, можно фиксировать желаемые значения чувствительности или специфичности или подбирать между ними компромисс.

Поэтому основной мерой качества модели является AUC — площадь под ROC-кривой (кривой в осях чувствительности от (1-специфичности)), которая не зависит от выбора порога классификации.

Для построения обучающей матрицы m_{dc} используется тот факт, что классы несбалансированны. Для каждой болезни количество больных превышает количество здоровых. Это оказывает прямое воздействие на специфичность и чувствительность, поэтому нужно балансировать веса классов. Первый столбец матрицы соответствует классу 0, второй — классу 1. Используется значение $m_{dc} = [d \in c]w(\bar{c})$, где $w(\bar{c})$ – количество элементов другого класса.

Следующие гипотезы использовались для построения наиболее точной модели:

- каждый класс описывается своим небольшим набором тем, 5-30 тем на класс
- также могуть быть фоновые темы, которые никак не связаны с классификациями, в каждом документе к фону относится 50%-90% слов.

Выделение фоновых тем и разрежевание остальных реализуется при помощи разреживающе-сглаживающего регуляризатора. Для этого использовались векторы параметров α_t и β_w. В качестве B указать индексы фоновых тем. Формулы М-шага имеют вид:

$$\theta_{td} \propto (n_{dt} + [t \in B]\alpha_t - [t \in S]\alpha_t + \tau_c m_{dt})_+$$

$$\varphi_{wt} \propto (n_{wt} + [t \in B]\beta_w - [t \in S]\beta_w)_+$$

Формула для θ включает в себя также регуляризатор классификации. Также из-за несбалансированности классов был использован частотный регуляризатор

$$\psi_{ct} = m_{ct} + \tau \frac{|D_c|\psi_{ct} m_t}{\sum_{s \in T} \psi_{cs} m(s)}$$

Начальное приближение в EM-алгоритме для матриц Φ и Θ выбиралось из равномерного распределения. Как показали эксперименты, элементы матрицы Ψ лучше всего инициализировать фиксированным значением 0.5, т.к. в модели всего два класса.

Основная задача состояла в том, чтобы научиться подбирать параметры модели, такие как степень разреженности матриц, количество

фоновых тем, параметры τ в регуляризаторах, чтобы получить высокое качество модели.

Для классификации тестовых выборок необходимо получить матрицу Θ^{test}. Для этого используется обычный EM-алгоритм с зафиксированной матрицей Φ^{train}.

3.4 Результаты эксперимента

Проводилось множество экспериментов с различными комбинациями регуляризаторов и параметров. В этом разделе приводятся выводы об использовании тех или иных параметров модели.

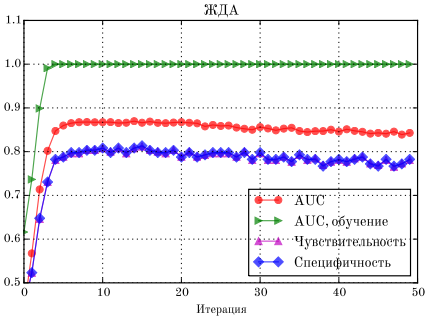
Эксперименты показали, что лучше всего выделять одну-две фоновых темы, подбирая α_t и β_w так, чтобы в кодограмме к фону относилось 60%–70% триграмм.

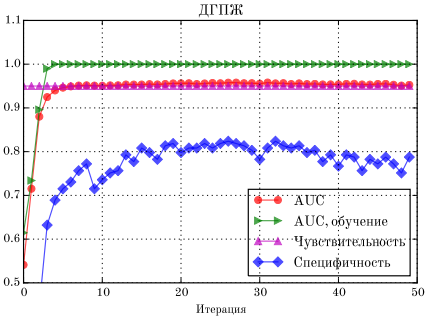
Коэффициент перед регуляризатором классификации необходимо выбирать как можно больше, т.к. он дает явную прибавку к качеству модели.

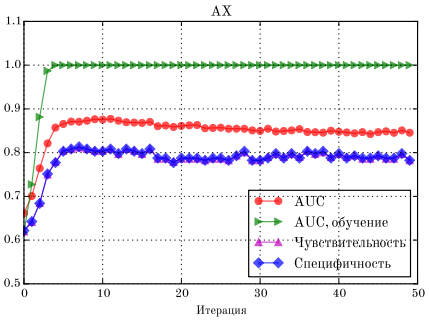
Несмотря на, казалось бы, очевидное и полезное свойство разреженности матриц, эксперименты показали, что лучшее качество классификации достигается при полном исключении разреживания из модели.

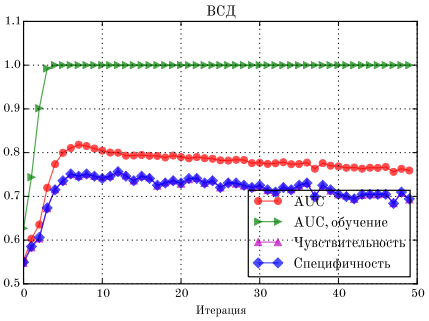
Как было замечено ранее, мы можем варьировать необходимый нам уровень чувствительности или специфичности, поэтому в экспериментах будем фиксировать высокую чувствительность на уровне 0.95, т.к. нам важно не пропустить болезнь. Однако не все болезни при столь высоком уровне чувствительности показывают достойное значение специфичности. Если в таких случаях специфичность падает ниже 0.6, будем выбирать порог так, чтобы две этих меры качества находились как можно ближе друг к другу.

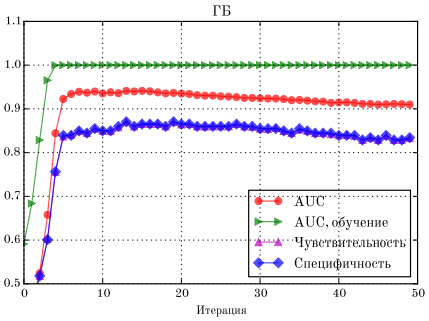
Далее приводятся графики зависимости мер качества модели от итераций EM-алгоритма на 10-блочной кросс-валидации для каждой из 18-ти болезней (рис 3.4).

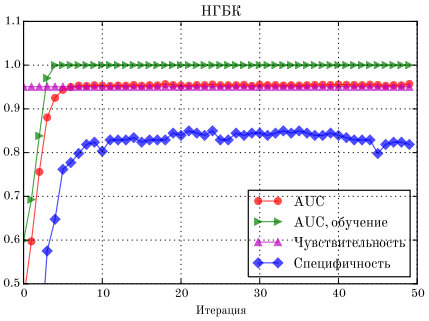

ЖДА
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

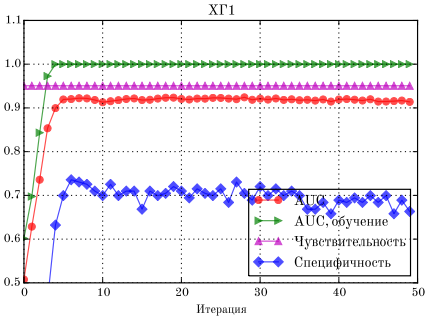

ДГПЖ
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

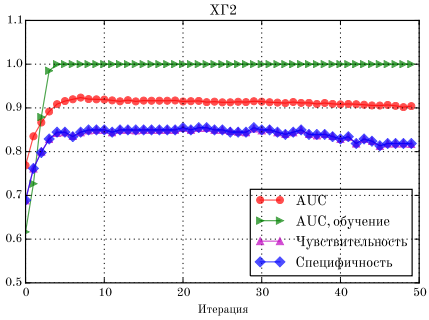

АХ
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

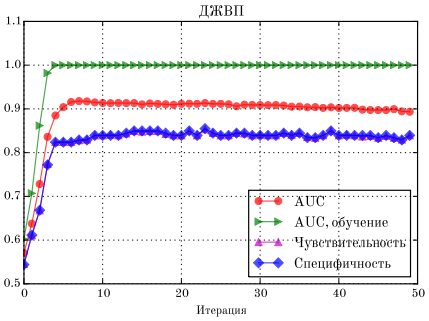

ВСД
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

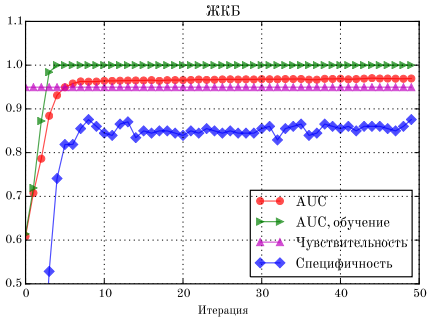

ГБ
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

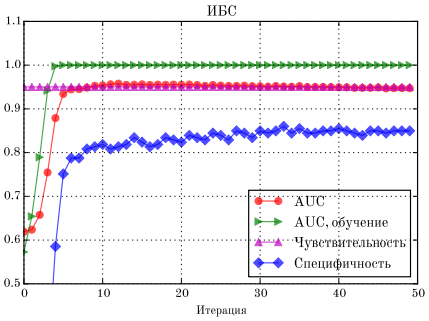

НГБК
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

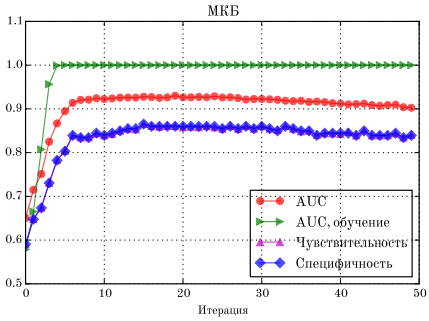

ХГ1
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

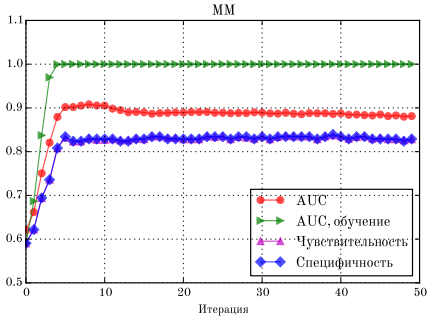

ХГ2
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

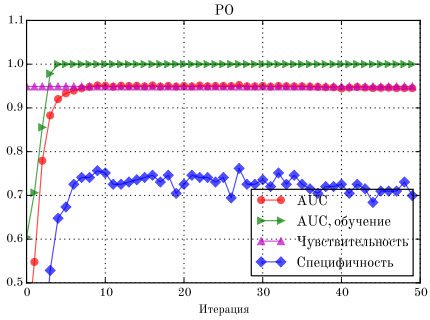

ДЖВП
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

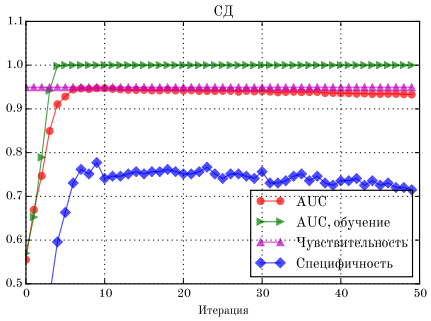

ЖКБ
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

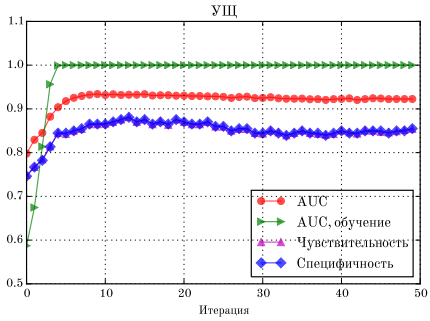

ИБС
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

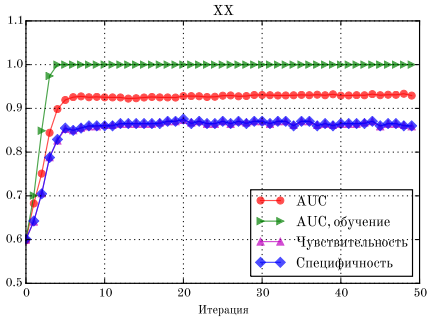

МКБ
AUC
AUC, обучение
Чувствительность
Специфичность
Итерация

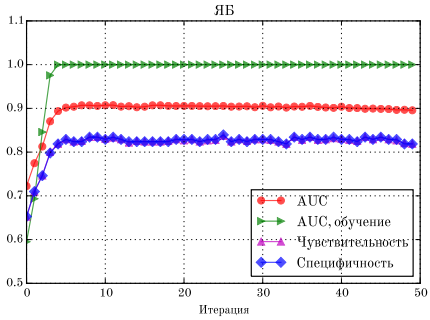












Из привденных графиков можно заметить, что некоторые болезни, например желчнокаменная болезнь (ЖКБ) или некроз головки бедренной кости (НГБК) показывают очень хорошее качество классификации,

а некоторые —, например, вегетососудистая дистония (ВСД) —очень плохое. Также заметно, что иногда алгоритм начинает сильно переобучаться (ВСД, ГБ) и качество классификации с ростом итераций уменьшается. Для таких случаев нужно контролировать количество итераций, применяя так называемый ранний останов.

Также полученный классификатор сравнивался с классическими алгоритмами машинного обучения, такими как линейный SVM и Random forest (100 деревьев). Стоит заметить, что все алгоритмы обучались и валидировались на одних и тех же разбиениях выборки. Также никакой обработки входных данных, как, например, отбор признаков не проводилось, что могло бы улучшить качество классификации. Результаты представлены в таблице 3.

Таблица 3: Сравнение тематической модели (ТМ) с другими алгоритмами классификации.

	ТМ	SVM	RF
АХ	**88.83**	77.08	86.50
ВСД	**82.51**	65.89	69.58
ГБ	**93.42**	81.11	81.58
ДГПЖ	**93.18**	84.71	90.72
ДЖВП	**91.70**	76.82	84.22
ЖДА	**85.34**	73.94	84.01
ЖКБ	**96.70**	89.66	94.89
ИБС	**97.00**	87.80	86.93
МКБ	**91.28**	80.98	84.67
ММ	**90.04**	78.55	80.72
НГБК	**95.67**	93.58	95.56
РО	**94.62**	84.09	90.45
СД	**94.47**	86.36	89.32
УЩ	**93.87**	80.17	86.13
ХГ1	**92.22**	85.16	91.05
ХГ2	**92.69**	80.96	85.13
ХХ	**92.46**	81.39	88.15
ЯБ	**89.73**	78.65	83.26

4 Заключение

В результате проведенных исследований была построена регуляризованная тематическая модель для классификации символьных последовательностей. В частности, полученный алгоритм был успешно применен для решения задачи классификации дискретизированных электрокардиосигналов. По итогам сравнения с классическими алгоритмами классификации показана целесообразность выделения нескольких тем (диагностических эталонов) для каждого заболевания.

Список литературы

[1] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[2] Tselykh V. R. Bunakov V. A. Uspenskiy V. M., Vorontsov K. V. Information function of the heart: Discrete and fuzzy encoding of the ecg-signal for multidisease diagnostic system. In *Series on Advances in Mathematics for Applied Sciences, World Scientific*, volume 86, pages 375–382, Singapore, 2015.

[3] K. V. Vorontsov. Additive regularization for topic models of text collections. In *Doklady Mathematics*, volume 89, pages 301–304. Pleiades Publishing, 2014.

[4] Успенский В. М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов. *Экономика и информатика*, 2008.