Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

Шаповалов Никита Анатольевич

Тематические модели для классификации символьных последовательностей в задачах биоинформатики и анализа биомедицинских сигналов

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Научный руководитель:

д.ф-м.н., доцент

К.В. Воронцов

Москва, 2016

Содержание

1 Введение **4**

1.1 Задачи классификации символьных последовательностей 4

1.2 Поиск участков ДНК, гиперчувствительных к ферментам 4

1.3 Тематическое моделирование в задачах биоинформатики 5

1.4 Постановка задачи 6

2 Методы классификации символьных последовательностей **6**

2.1 Линейные модели классификации 7

2.2 Наивный байесовский классификатор 7

2.3 Логистическая регрессия 8

2.4 Метод опорных векторов 9

2.5 Случайный лес 9

2.6 Тематические модели классификации 9

2.7 Оценивание качества бинарной классификации 12

3 Тематическая модель классификации участков ДНК **12**

3.1 Подготовка данных 12

3.2 Зависимость качества классификации ТМ от параметров 13

3.3 Оптимизация параметров тематической модели классификации 15

3.4 Сравнение с другими моделями классификации 16

4 Результаты, выносимые на защиту **18**

Список литературы **19**

Аннотация

Методы классификации символьных последовательностей широко применяются в компьютерной лингвистике для анализа текстов естественного языка и в биоинформатике для анализа нуклеотидных и аминокислотных последовательностей. Вероятностные тематические модели изначально разрабатывались как инструмент кластеризации и выявления латентной семантической структуры коллекций текстовых документов. Затем выяснилось, что знание этой структуры улучшает качество классификации текстовых документов. В последнее время тематическое моделирование всё чаще применяется для анализа и классификации символьных последовательностей неязыковой природы; расширяется спектр его приложений в области биоинформатики и анализа дискретных биомедицинских сигналов.

В данной работе рассматривается задача классификации участков нуклеотидных последовательностей, гиперчувствительных к ферментам. Целью работы является проверка возможности применения математического аппарата аддитивной регуляризации тематических моделей. Для этого тематические модели сравниваются с другими методами классификации символьных последовательностей, в том числе с линейными моделями и со случайным лесом. Эксперименты показывают, что путём подбора управляющих параметров тематической модели возможно добиваться высокого качества классификации по критерию AUC.

1 Введение

1.1 Задачи классификации символьных последовательностей

В последнее время кроме анализа текстов стали появляться задачи обработки символьных последовательностей. Большой категорией таких последовательностей являются биомедицинские сигналы, классификация которых может представлять огромную теоретическую и практическую ценность. Например, анализ сигналов ЭКГ [11] позволяет выявлять опасные патологии после обычного обследования. В частности, получают развитие методы символьной динамики.

Для решения задач классификации символьных последовательностей произвольной длины чаще всего переходят к признаковому описанию фиксированного размера. Например, подсчитывают частоты встречаемости всех *n-грамм* – последовательностей длины n, состоящих из символов того же алфавита. Для анализа некоторых биомедицинских сигналов оказываются эффективными линейные модели, которые анализируют вектор частот n-грамм [1]. Однако линейные модели применимы только для линейно разделимых классов, что может снижать успешность использования.

1.2 Поиск участков ДНК, гиперчувствительных к ферментам

Одной из важных задач, возникающих в биоинформатике, является поиск участков ДНК, чувствительных к ферментам. Известно, что достаточно хорошо данную задачу решают линейные методы классификации, в которых предполагается, что классы линейно разделимы в пространстве признаков.

Установление нуклеотидных последовательностей более 1000 геномов и, прежде всего, генома человека, обусловило необходимость функциональной аннотации генома. Данная задача только лишь отчасти может быть решена экспериментально и разработка и применение надежных вычислительных методов к задаче аннотации генома чрезвычайно востребованы.

Глобальная задача аннотации генома состоит из множества вычислительных задач, связанных с теми или иными разновидностями биологических ролей составляющей геном ДНК. В частности, одной из таких задач является необходимость определения участков в последовательности геномной ДНК, которые характеризуются повышенной чувствительностью к ДНК-трансформирующим ферментам (прежде всего, к ДНКазе I). Распознавание таких участков в последовательности ДНК может иметь важное значение для установления нук-

леотидных вариаций генома, которые взаимосвязаны с повышенной склонностью пациентов к развитию тех или иных заболеваний. Гиперчувствительность определенных отрезков ДНК к ДНКазе I является весьма точным маркером регуляторных элементов ДНК, отвечающих за инициацию транскрипции ДНК в РНК [10].

В рамках машинного обучения поиск гиперчувствительных участков нуклеотидной последовательности можно свести к решению задачи классификации, когда по имеющемуся признаковому описанию учатска необходимо понять, является ли он чувствительным к данному ферменту (положительный класс) или нет (отрицательный класс).

1.3 Тематическое моделирование в задачах биоинформатики

Тематическое моделирование является мощным инструментом для статистического анализа текстов. Основной идеей тематического моделирования является выявление скрытых структур в тексте — тем, смеси которых составляют документы. Каждая тема представляет собой дискретное вероятностное распределение на множестве слов. Тематическая модель классификации оперирует не только со словами, но и с метками классов, позволяя для каждой темы определить, насколько ярко каждая тема принадлежит тому или иному классу.

Тематические модели классификации могут описывать гораздо более разнообразные структуры классов, чем линейные модели. В линейных моделях классификации все ограничивается разделяющей гиперплоскостью, тогда как в тематических моделях темы могут описывать кластеры отдельных классов, что может повышать качество классификации [8].

Тематические модели оказались полезны не только для анализа текстов, но и в тех областях, где высока вероятность присутствия скрытых структур.

В задаче анализа геномных аннотаций [7] необходимо понять, какие белки могут взаимодействовать между собой. Каждая аннотация описывает только гены, кодирующие белки, и для каждого такого гена определяет, какую функцию может выполнять кодируемый им белок. В статье авторы рассматривают каждую аннотацию как документ, в котором словами являются не гены, а функции, которые выполняют кодируемые ими белки. Темам они придают смысл биологических процессов. Для нахождения тем и разбиения документов используется модель латентного размещения Дирихле (LDA)[2].

В задаче поиска смежных экспериментов с ДНК-микрочипами [4] ставится задача определения экспериментов, в которых протекали похожие биологические процессы. В качестве

исходных данных выступает набор экспериментов с ДНК-микрочипами по оцениванию экспрессии генов человека в различных внешних условиях. Для каждого микрочипа известны внешние условия, при которых проводился соответствующий эксперимент, В каждом эксперименте в разных микрочипах находятся гены людей с разными фенотипами. Для получения информации из эксперимента берутся два подмножества микрочипов с разными фенотипами и определяется, экспрессия каких генов сильно отличается в этих наборах. Таким способом, для каждого эксперимента есть подмножество генов, полученное описанным выше способом. Используется тематическое моделирование: документами являются эксперименты, слова – гены, в документе есть слова только из полученного подмножества. Для оценивания похожести экспериментов считается вероятность того, что слова из одного эксперимента были получены на основе тематического профиля другого документа.

Наконец,тематическое моделирование применяется для анализа третичной структуры белков [9]: необходимо по набору вторичных структур, составляющих белок, научиться определять, какие белки имеют похожие третичные структуры. Каждое описание третичной структуры белка рассматривается как документ, состоящий из слов. Словарем является множество всех вторичных структур белка. После обработки белковых структур каждый белок отождествляется со своим дискретным вероятностным распределением на темах в виде вектора фиксированной длины. Для определения сходства белков (т.е. документов) применяются косинус-мера и дивергенция Кульбака-Лейблера (en. *KL-divergence*).

1.4 Постановка задачи

Цель работы – проверить гипотезу, что тематическая модель классификации подходит для поиска гиперчувствительных участков ДНК, для этого предлагается сравнить тематическую модель классификации с другими методами.

2 Методы классификации символьных последовательностей

Пусть X – множество описаний объектов, Y – множество меток классов. Предполагается, что существует неизвестная зависимость – отображение $y : X \to Y$, значения которой известны только для объектов обучающей выборки $X^N = \{(x_1, y_1) \dots (x_N, y_N)\}$. В задаче

классификации требуется по имеющейся обучающей выборке построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

Бинарная классификация является частным случаем, когда множество Y состоит из двух элементов. Без ограничения общности можно считать, что $Y = \{-1, +1\}$.

2.1 Линейные модели классификации

В основе *линейных моделей классификации* [6] лежит предположение о линейной разделимости объектов различных классов. В случае бинарной классификации в качестве разделяющей поверхности выступает гиперплоскость.

Формально говоря, пусть каждый объект описывается набором числовых признаков $x = (x_1, \ldots, x_D) \in \mathbb{R}^D$. Пусть множество меток класса $Y - \{0, 1\}$. Линейный классификатор $a : X \to Y$ имеет вид:

$$a(x, w) = \text{sign}\left(\sum_{j=1}^{D} x^j w^j - b\right) = \text{sign}\left(\langle x, w\rangle - b\right), \tag{1}$$

где w^j – вес j-го признака, b – порог принятия решения.

Различные линейные классификаторы различаются по методу обучения параметров $(w^1, \ldots, w^D, b)$.

2.2 Наивный байесовский классификатор

Согласно байесовской теории классификации, минимальной ошибкой классификации обладает *байесовский классификатор*, который относит объект x к тому классу y, чья *апостериорная вероятность* $P(y|x)$ максимальна:

$$a(x) = \arg\max_{y \in Y} P(y|x) = \arg\max_{y \in Y} P(x|y)P(y), \tag{2}$$

где $P(y)$ – *априорная вероятность* класса y, $p(x|y)$ – *условная плотность распределения* класса y. В задачах обучения классификации плотность $p(x, y)$ не известна; известна лишь конечная обучающая выборка $(x_i, y_i) \in X \times Y$, $i = 1, \ldots, N$

Предполагается, что объекты обучающей выборки получены в результате случайных и независимых испытаний из распределения с плотностью $p(x, y)$.

Пусть объекты описываются D-мерными векторами признаков $x = (x^1, \ldots, x^D)$, где каждый признак может принимать целые неотрицательные значения: $x^j \in \mathbb{Z}_+$. В случае наивного байесовского классификатора делается довольно сильное предположение, что признаки являются статистически независимыми случайными величинами. В случае наивного байесовского классификатора предполагается, что при фиксированной метке класса каждый признак имеет пуассоновское распределение. Объединяя эти предположения, получаем, что условная вероятность объекта имеет вид:

$$p(x|y) = \prod_{j=1}^{D} p(x^j|y) = \prod_{j=1}^{D} Poiss(x^j|\mu_y^j) = \prod_{j=1}^{D} \frac{(\mu_y^j)^{x^j}}{x^j!} \exp(-\mu_y^j) \tag{3}$$

Параметры настраиваются путем максимизации логарифма правдоподобия:

$$\mathcal{L}(X, \boldsymbol{\mu}) = \sum_{i=1}^{N} \log p(x_i|y_i) = \sum_{i=1}^{N} \sum_{j=1} \log p(x_i^j|\mu_{y_i}^j) \to \max_{\boldsymbol{\mu}} \tag{4}$$

Нетрудно показать, что полученные значения параметров являются выборочными средними признаков по объектам соответствующего класса: $\mu_y^j = \frac{\sum\limits_{i=1}^{N} [y_i=y] x_i^j}{\sum\limits_{i=1}^{N} [y_i=y]}$. Также полученный классификатор является линейным:

$$\begin{aligned} w^j &= \ln \mu_{+1}^j - \ln \mu_{-1}^j, \\ b &= \ln P(y=-1) - \ln P(y=+1). \end{aligned} \tag{5}$$

2.3 Логистическая регрессия

В логистической регрессии для поиска весов $w^1, \ldots, w^D, b$ решается следующая оптимизационная задача:

$$\mathcal{Q}(w, b) = \sum_{i=1}^{N} \ln(1 + \exp(-y_i(\langle x_i, w\rangle - b))) \to \min_{w,b} \tag{6}$$

Для улучшения обобщающей способности обычно применяется l2-регуляризация:

$$\mathcal{Q}(w, b) = \sum_{i=1}^{N} \ln(1 + \exp(-y_i(\langle x_i, w\rangle - b))) + \frac{\lambda}{2}(\|w\|_2^2 + b^2) \to \min_{w,b}, \tag{7}$$

где $\lambda > 0$ – коэффициент регуляризации – настраиваемый гиперпараметр.

2.4 Метод опорных векторов

В методе опорных векторов настройка весов $w^1, \ldots, w^D, b$ происходит путем решения оптимизационной задачи:

$$\begin{cases} \frac{1}{2}\|w\|_2^2 + C\sum\limits_{i=1}^{N}\xi_i \to \min\limits_{w,b,\xi}, \\ y_i(\langle x_i, w\rangle - b) \geq 1 - \xi_i,\ i = 1, \ldots, N, \\ \xi_i \geq 0, \end{cases} \tag{8}$$

где $C > 0$ – коэффициент регуляризации – настраиваемый гиперпараметр.

2.5 Случайный лес

Случайный лес[3] представляет собой композицию решающих деревьев на основе *бэггинга*: каждое решающее дерево строится на основе случайной подвыборки, полученной выбором l объектов с возвращением – *бутстреппингом*. Кроме того, при создании очередного узла решающего дерева выбор признака, по которому будет производиться разбиение, происходит не из всего множества признаков, а из случайного подмножества.

Обозначим за $q_m(x)$ долю решающих деревьев, которые относят объект x к классу $+1$. Тогда классификатор имеет следующий вид:

$$a(x) = [q_m(x) \geq c]\,, \tag{9}$$

где c – порог принятия решения (чаще всего $c = 1/2$).

2.6 Тематические модели классификации

Пусть D — коллекция текстовых документов, W^1 — множество употребляемых в них слов. Документ может содержать не только слова, но и метаданные различных типов (*модальностей*). Каждая модальность определяется конечным набором элементов $W^m, m = 1, \ldots, M$.

Любой документ $d \in D$ представляет собой последовательность токенов различных модальностей: $(w_1, \ldots, w_{n_d}), w_i \in W$, где $W = W^1 \sqcup \ldots \sqcup W^m$. Предполагается гипотеза *мешка*

слов (порядок элементов в документе не важен), благодаря которой документ отождествляется с набором $\{n_{dw} \,|\, w \in W\}$, представляющим собой число вхождений каждого токена w в документ d. Также предполагается, что существует конечное множество тем T, и каждое появление токена w в документе d связано с темой $t \in T$, которая заранее не известна. Вкупе с этими предположениями, коллекция документов рассматривается как набор независимых троек $(w_i, d_i, t_i), i = 1, \ldots, n$, полученных из дискретного распределения $p(w, d, t)$ на вероятностном пространстве $D \times W \times T$.

Запишем вероятностную тематическую модель (BTM) для каждой модальности [12, 13]:

$$p(w|d) = \sum_{t \in T} p_m(w|t)p(t|d) = \sum_{t \in T} \phi^m_{wt}\theta_{td}, d \in D, w \in W^m, m = 1, \ldots, M.$$

Параметры $\theta_{td} = p(t|d)$ и $\phi^m_{wt} = p_m(w|t)$ образуют матрицы $\Theta = (\theta_{td})_{T \times D}$ – дискретные распределения тем для документов, и $\Phi^m = (\phi_{wt})_{W^m \times T}$ – дискретные распределения токенов каждой модальности для тем. Из этого следует, что эти матрицы являются *стохастическими* (каждый столбец представляет собой дискретное распределение). Обозначим блочную матрицу-столбец $(\Phi^1, \ldots \Phi^m)^T$ за Φ.

Чтобы найти неизвестные элементы матриц Φ^m, Θ по наблюдаемой коллекции документов, максимизируем логарифм правдоподобия для каждой модальности:

$$\mathcal{L}_m(\Phi^m, \Theta) = \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln p_m(w|d) \to \max_{\Phi^m, \Theta}$$

В рамках аддитивной регуляризации тематических моделей (APTM) [12, 13], данная многокритериальная задача преобразуется в задачу условной оптимизации:

$$\sum_{m=1}^{M} \tau_m \mathcal{L}_m(\Phi^m, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \tag{10}$$

$$\sum_{w \in W^m} \phi_{wt} = 1, \phi_{wt} \geq 0; \tag{11}$$

$$\sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0, \tag{12}$$

где $R(\Phi, \Theta)$ – дополнительный критерий-регуляризатор, учитывающий специфику задачи и знания предметной области.

Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Известно [12, 13], что в таком случае локальный максимум задачи (10)-(12) удовлетворяет следующей системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d, w)$:

$$p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right); \tag{13}$$

$$\phi_{wt} = \underset{w\in W^{m(w)}}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \qquad n_{wt} = \sum_{d\in D}\tau_{m(w)}n_{dw}p_{tdw}; \tag{14}$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_{w\in d}\tau_{m(w)}n_{dw}p_{tdw}; \tag{15}$$

где оператор $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_s,0\}}$; $m(w)$ — модальность токена w: $w \in W^{m(w)}$.

Решение системы уравнений (13)-(15) с помощью метода простой итерации эквивалентно EM-алгоритму, где на E-шаге мы пересчитываем значения p_{tdw} согласно (13), а на M-шаге вычисляем ϕ_{wt} и θ_{td} на основе (14)-(15).

Тематические модели классификации являются частным случаем мультимодальных тематических моделей, в которых каждый документ может содержать метки классов, относящиеся к отдельной модальности C. Для удобства часть матрицы Φ, соответствующую модальности классов C обозначают за Ψ.

Обучение тематической модели Пусть у нас имеется коллекция документов D_{train}, в которой каждый документ содержит слова $w \in W$, а также метки классов $c \in C$. Обучение ТМ состоит в решении задачи (10)-(12) с помощью EM-алгоритма.

Классификация тестовых документов Пусть у нас имеется коллекция документов D_{test}, в которой каждый документ содержит только слова $w \in W$. Тематическая модель классификации позволяет на основе этой информации получить $p(c|d), c \in C, d \in D_{test}$:

- Неизвестные θ_{td} находятся путем решения задачи условной оптимизации (10)-(12) при фиксированной матрице Φ;
- Вычисляются $p(c|d) = \sum_{t\in T}\psi_{ct}\theta_{td}$.

Построение мультимодальных тематических моделей осуществлялось с использованием библиотеки тематического моделирования BigARTM[1] [12].

[1] `http://bigartm.org`

2.7 Оценивание качества бинарной классификации

При решении задач бинарной классификации основными критериями качества алгоритма $a(x)$ являются:

- Чувствительность: $\dfrac{\sum\limits_{k=1}^{N}[y_k = +1]\,[a(x_k) = +1]}{\sum\limits_{k=1}^{N}[y_k = +1]}$;
- Специфичность: $\dfrac{\sum\limits_{k=1}^{N}[y_k = -1]\,[a(x_k) = -1]}{\sum\limits_{k=1}^{N}[y_k = -1]}$;
- Площадь под ROC-кривой (*Area Under ROC, AUROC*). ROC-кривая – множество достигаемых пар (1-специфичность, чувствительность) при всевозможных значениях порога классификации.

Поскольку последний критерий качества не зависит от порога классификации, то для сравнения моделей будем использовать именно его.

3 Тематическая модель классификации участков ДНК

3.1 Подготовка данных

В настоящем исследовании использовались следующие данные:

- Обобщённая ДНК человека, разбитая по хромосомам. Каждая хромосома представлена в виде строки из символов *A, G, C, T, N*(означает, что нуклеотидное основание, стоящее на данной позиции, не удалось определить);
- Данные экспериментов по локализации гиперчувствительных к ДНКазе-I участков в клетках A549. Про каждый имеющийся гиперчувствительный участок известно, в какой хромосоме он находится, а также позиции в этой хромосоме. Отдельно известно, что каждый гиперчувствительный участок имеет длину 150 нуклеотидных пар.

Таким образом, для каждого гиперчувствительного участка мы знаем последовательность нуклеотидов, из которых он состоит.

Получение участков, не являющихся гиперчувствительными Особенность данных такова, что у нас имеются только участки хромосом, гиперчувствительные к ДНКазе-I, но при этом мы не имеем никакой информации об участках, которые заведомо таковыми не являются.

В [5] предлагается исследовать способность алгоритма отличать гиперчувствительные участки от случайно выбранных участков последовательности ДНК. Для того, чтобы не затрагивать чувствительные участки, был предложен и реализован следующий подход:

- В каждой хромосоме удаляем гиперчувствительные участки, а также участки с неопределенными нуклеотидами. После этого каждая хромосома разбивается на фрагменты.
- Оставшиеся фрагменты разбиваются на части примерно одинаковой длины, из которых затем случайным образом выбирается непрерывный участок с длиной, равной длине имеющихся объектов (т.е. 150 нуклеотидных пар). Для получения сопоставимого по числу объектов отрицательного класса длина части, на которую бьются фрагменты, была определена экспериментально.

Векторизация символьной последовательности Пусть у нас есть последовательность $\{l_k\}$ символов конечного алфавита $\mathcal{A}$. *n-граммой* называется последовательность $l_t, l_{t+1}, \ldots, l_{t+n-1}$ n подряд идущих элементов последовательности L. Множество всех возможных n-грамм $W = \mathcal{A}^n$ содержит $|\mathcal{A}|^n$ элементов. В нашем случае $|\mathcal{A}| = 4$.

Процесс *разбиения* символьной последовательности на *n-граммы* состоит в выделении и упорядочении всех n-грамм данной символьной последовательности. При этом каждой последовательности символов, имеющей длину хотя бы n, соответствует ровно одна последовательность n-грамм, и наоборот.

Частотой n-граммы $s = \{s_1, s_2, \ldots, s_n\}$ будем называть число её вхождений в последовательность L и обозначать $n_s(L)$.

При фиксированном n *векторизацией* символьной последовательности L будем называть процесс получения вектора частот n-грамм $(n_s(L), s \in W)$. Заметим, что для нуклеотидных последовательностей данный вектор имеет длину 4^n.

3.2 Зависимость качества классификации ТМ от параметров

Целью данного эксперимента было исследование зависимости качества классификации ТМ от признакового описания участков, а также от параметров модели.

	$\tau = 30$	$\tau = 100$	$\tau = 300$	$\tau = 1000$
$\|T\| = 10$	0.8024	0.8214	0.8333	0.8381
$\|T\| = 20$	0.8017	0.8228	0.8443	0.8417
$\|T\| = 50$	0.8047	0.8326	0.8474	0.8478
$\|T\| = 100$	0.8048	0.8357	0.8502	**0.8520**

Таблица 1: Значение AUROC для TM классификации, 5-граммы

	$\tau = 30$	$\tau = 100$	$\tau = 300$	$\tau = 1000$
$\|T\| = 10$	0.8248	0.8467	0.8578	0.8559
$\|T\| = 20$	0.8173	0.8613	0.8636	0.8642
$\|T\| = 50$	0.8162	0.8577	0.8657	0.8652
$\|T\| = 100$	0.8223	0.8578	**0.8684**	**0.8683**

Таблица 2: Значение AUROC для TM классификации, 6-граммы

Параметры эксперимента Признаковое описание участков нуклеотидной последовательности определялось вектором частот n-грамм, n варьировалось от 5 до 7 включительно.

Параметры мультимодальной TM:

- Число тем $|T| \in \{10, 20, 50, 100\}$;
- Коэффициент модальности классов $\tau \in \{30, 100, 300, 1000\}$.

Критерий качества оценивался по 5-блочной кросс-валидации. При фиксированной обучающей и тестовой выборке производилось 35 проходов по коллекции для построения модели из 10 случайных начальных приближений. Значение AUROC вычислялось на обучении и контроле после каждого прохода по коллекции для каждого случайного приближения, после чего бралось максимальное значение.

Результаты эксперимента и выводы Результаты показаны в табл. 1–табл. 3.

По результатам эксперимента можно сделать следующие выводы:

- Значение критерия качества увеличивается как при увеличении параметра n, отвечающего за признаковое описание участков, так и при увеличении параметров TM классификации в рассматриваемом диапазоне параметров. либо критерий качества может незначительно уменьшаться;

	$\tau = 30$	$\tau = 100$	$\tau = 300$	$\tau = 1000$
$\|T\| = 10$	0.8327	0.8622	0.8736	0.8724
$\|T\| = 20$	0.8364	0.8695	0.8783	0.8791
$\|T\| = 50$	0.8395	0.8722	0.8807	0.8815
$\|T\| = 100$	0.8392	0.8735	0.8824	**0.8831**

Таблица 3: Значение AUROC для TM классификации, 7-граммы

- При увеличении значений параметров модели их влияние на критерий качества ослабевает.
- Можно предположить, что при больших значениях параметров TM классификации качество перестает увеличиваться.

Последнее предположение было решено проверить экспериментально.

3.3 Оптимизация параметров тематической модели классификации

Алгоритм для оптимизации параметров тематической модели основан на гипотезе, выдвинутой по результатам предыдущего эксперимента: при фиксировании всех параметров, кроме одного, качество классификации AUROC образует унимодальную функцию от этого параметра.

Алгоритм представляет из себя покоординатный подъем: на каждом шаге мы фиксируем либо число тем $|T|$, либо коэффициент модальности классов τ_c, и далее оптимизируем качество классификации по другому параметру. Поскольку предполагается, что функция является унимодальной, то оптимизация осуществляется с помощью троичного поиска. Для параметра, определяющего число тем $|T|$, оптимизация производилась по всем числам, кратным 10. Для параметра, определяющего коэффициент модальности классов τ_c, оптимизация производилась по всем натуральным числам.

Данный алгоритм был реализован и экспериментально проверен: была произведена настройка параметров описанным выше способом, начальное приближение определялось параметрами $|T| = 30$ и $\tau_c = 30$. В качестве данных использовались векторные частоты 7-грамм.

Результаты эксперимента и выводы В ходе работы алгоритма были выявлены оптимальные значения параметров $\tau_c = 1153, |T| = 120$.

$\|T\| = 10$	$\|T\| = 40$	$\|T\| = 100$	$\|T\| = 120$	$\|T\| = 150$	$\|T\| = 200$
0.8740	0.8799	0.8842	**0.8850**	0.8845	0.8840

Таблица 4: Зависимость AUROC от числа тем $|T|$, $\tau_c = 1150$

$\tau_c = 100$	$\tau_c = 600$	$\tau_c = 900$	$\tau_c = 1153$	$\tau_c = 1500$	$\tau_c = 1800$
0.8751	0.8837	0.8846	**0.8850**	0.8848	0.8845

Таблица 5: Зависимость AUROC от коэффициента модальности классов τ_c, $|T| = 120$

В табл. 4 и табл. 5 показаны значения критерия качества для параметров ТМ при фиксировании другого параметра.

По результатам данного эксперимента можно сделать выводы о том, что гипотеза не отвергается.

3.4 Сравнение с другими моделями классификации

Цель данного эксперимента – проверить, насколько хорошо мультимодальные тематические модели решают данную задачу классификации по сравнению с другими алгоритмами машинного обучения. Для сравнения использовались l_2-регуляризованная логистическая регрессия, метод опорных векторов, случайный лес и наивный пуассоновский байес.

Параметры эксперимента В качестве признаковых описаний участок нуклеотидных последовательностей использовались векторы частот 7-грамм. Таким образом, число признаков составляло $4^7 = 16384$.

Выборка была разделена на обучение и контроль в отношении 2 к 1. Параметры логистической регрессии и SVM – регуляризационные коэффициенты – настраивались путем 5-блочной кросс-валидации на обучающей выборке по сетке.

Параметры для обучения случайного леса брались таким образом, чтобы суммарное процессорное время не превышало одного часа: строилась композиция из 100 решающих деревьев, для построения одного узла использовалось логарифмическое число признаков ($O(\log_2 D)$).

Параметры тематической модели классификации: число тем $|T| = 100$, коэффициент при модальности классов $\tau_c = 1000$, коэффициент при модальности слов $\tau_w = 1$. При обучении делалось 15 шагов ЕМ-алгоритма. При классификации делался 1 шаг ЕМ-алгоритма.

Результаты эксперимента и выводы Результаты показаны на рис. 1 и рис. 2.

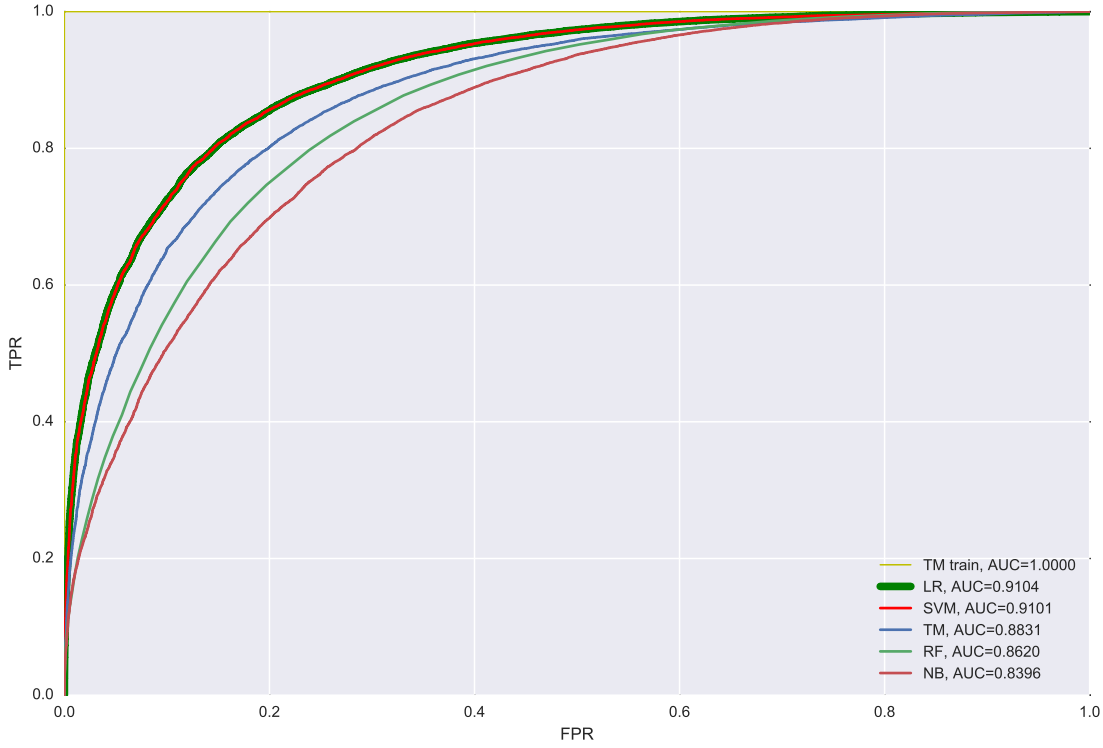

Рис. 1: ROC-кривые для различных алгоритмов классификации

В целом видно, что показатели качества тематической модели классификации лучше, чем у случайного леса или наивного байеса, но хуже, чем у логистической регрессии и метода опорных векторов. Однако при уменьшении специфичности до 0.3 тематическая модель классификации имеет меньшую чувствительность, чем у всех остальных сравниваемых алгоритмов. При этом тематическая модель всегда идеально классифицирует обучающую выборку.

Первая мысль, которая приходит в голову – переобучение. В самом деле, в тематической модели классификации при использовании в качестве слов 7-грамм на 100 темах мы имеем порядка 1.5 млн параметров, которые надо настраивать, тогда как при обучении мы имеем в 10 раз меньше документов.

Однако дополнительные исследования показали, что данная тенденция наблюдается вне зависимости от числа тем и числа признаков. Например, при настройке тематической модели классификации по векторам частот 5-грамм на 10 темах мы опять получаем, что на обучающей выборке идеальная классификация.

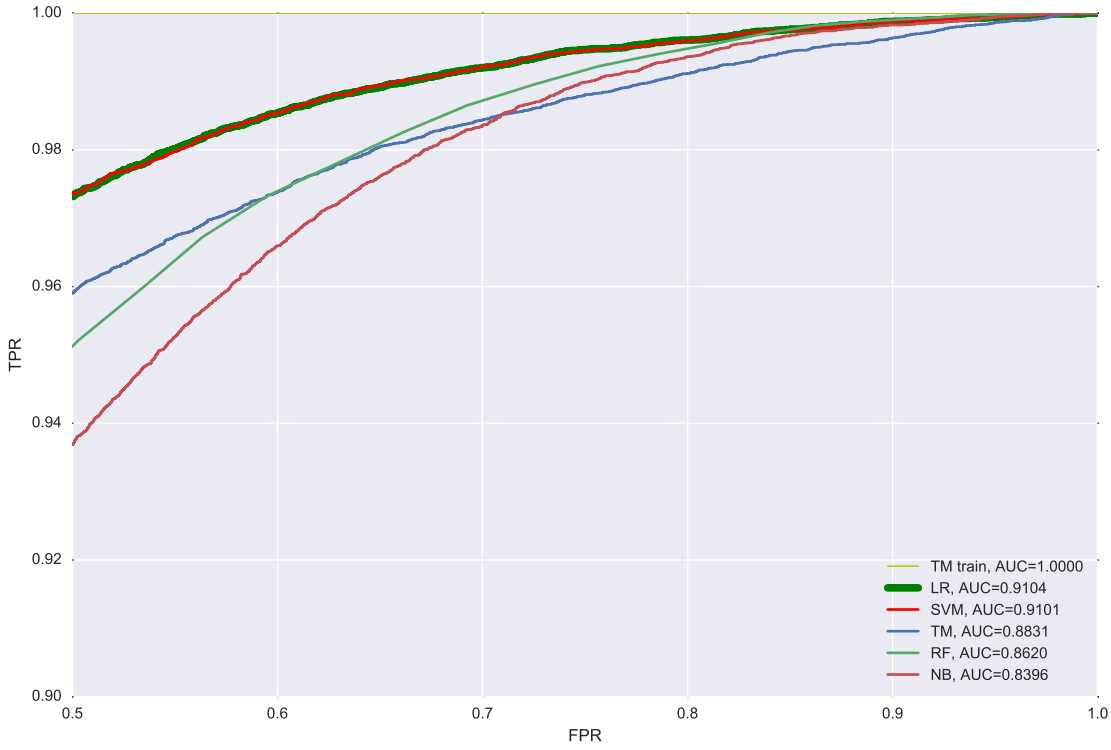

Рис. 2: ROC-кривые для различных алгоритмов классификации, $FPR \geq 0.5$

Единственный параметр, который действительно влияет – это коэффициент при модальности классов τ_c: при $\tau_c \gg 1$ мы получаем идеальную классификацию на обучающей выборке, при $\tau_c \approx 1$ мы получаем что качество классификации на обучающей выборке и на тестовой выборке коррелируют, однако оно существенно меньше, чем качество на тестовой выборке при $\tau_c = 1000$.

4 Результаты, выносимые на защиту

- Предложена тематическая модель классификации для поиска участков ДНК, гиперчувствительных к ферментам. Данная модель может также применяться для классификации символьных последовательностей любой природы.

- В экспериментах показано, что тематическая модель классификации немного уступает линейным моделям, опережая случайный лес и наивный байесовский классификатор.

Список литературы

[1] *Annala, M.* A linear model for transcription factor binding affinity prediction in protein binding microarrays / M. Annala, K. Laurila, H. Lähdesmäki, M. Nykter // *PLoS ONE.* — 05 2011. — Vol. 6, no. 5. — Pp. 1–13.

[2] *Blei, D. M.* Latent dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // *J. Mach. Learn. Res.* — mar 2003. — Vol. 3. — Pp. 993–1022.

[3] *Breiman, L.* Random Forests / L. Breiman // *Machine Learning.* — 2001. — Vol. 45. — Pp. 5–32.

[4] *Caldas, J.* Probabilistic retrieval and visualization of biologically relevant microarray experiments / J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, S. Kaski // *Bioinformatics.* — 2009. — Vol. 25, no. 12. — Pp. i145–i153.

[5] Evaluation of methods for modeling transcription factor sequence specificity / M. T. Weirauch, A. Cote, R. Norel, M. Annala, Y. Zhao, T. R. Riley, J. Saez-Rodriguez et al. // *Nature biotechnology.* — 2013. — Vol. 31, no. 2. — Pp. 126–134.

[6] *Hastie, T.* The Elements of Statistical Learning, 2nd edition / T. Hastie, R. Tibshirani, J. Friedman. — Springer, 2009. — 533 pp.

[7] *Konietzny, S.* Inferring functional modules of protein families with probabilistic topic models / S. Konietzny, L. Dietz, A. McHardy // *BMC Bioinformatics.* — 2011. — Vol. 12, no. 1. — P. 141.

[8] *Rubin, T. N.* Statistical topic models for multi-label document classification / T. N. Rubin, A. Chambers, P. Smyth, M. Steyvers // *Machine Learning.* — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[9] *Shivashankar, S.* Multi-view methods for protein structure comparison using latent dirichlet allocation / S. Shivashankar, S. Srivathsan, B. Ravindran, A. V. Tendulkar // *Bioinformatics.* — jul 2011. — Vol. 27, no. 13. — Pp. i61–i68.

[10] Systematic localization of common disease-associated variation in regulatory DNA / M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds et al. // *Science.* — 2012. — Vol. 337, no. 6099. — Pp. 1190–1195.

[11] *Uspenskiy, V.* Discrete and Fuzzy Encoding of the ECG-Signal for Multidisease Diagnostic System / V. Uspenskiy, K. Vorontsov, V. Tselykh, V. Bunakov // Advanced Mathematical and Computational Tools in Metrology and Testing X. — 2015. — Pp. 377–384.

[12] *Vorontsov, K.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. // AIST / Ed. by M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets. — Vol. 542 of *Communications in Computer and Information Science*. — Springer, 2015. — Pp. 370–381.

[13] *Vorontsov, K.* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization / K. Vorontsov, A. Potapenko // Analysis of Images, Social Networks and Texts / edited byD. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, R. E. Yavorsky. — Springer International Publishing, 2014. — Vol. 436 of *Communications in Computer and Information Science*. — Pp. 29–46.