На правах рукописи

Офицеров Евгений Петрович

Обработка символьных последовательностей методами машинного обучения на основе дифференцируемого выравнивания

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Тула — 2019

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Тульский государственный университет».

Научный руководитель: доктор физико–математических наук
Горбачев Дмитрий Викторович

Официальные оппоненты: **Бабенко Александр Григорьевич,**
доктор физико-математических наук, профессор,
Институт математики и механики им. Н.Н. Красовского Уральского отделения Российской академии наук, г. Екатеринбург,
заведующий отделом аппроксимации и приложений

Касьянов Артем Сергеевич,
кандидат физико-математических наук,
Институт проблем передачи информации им. А.А. Харкевича, г. Москва,
старший научный сотрудник

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого Президента России Б. Н. Ельцина», г. Екатеринбург

Защита состоится 11 февраля 2020 г. в 14 часов на заседании диссертационного совета Д 212.271.05, созданного на базе Тульского государственного университета по адресу: 300012, г. Тула, пр. Ленина, 92 (12-105).

С диссертацией можно ознакомиться в библиотеке Тульского государственного университета и на сайте http://tsu.tula.ru/science/dissertation/diss-212-271-05/Ofitserov_EP/.

Автореферат разослан 6 декабря 2019 года.

Ученый секретарь
диссертационного совета Соколова Марина Юрьевна

Общая характеристика работы

Актуальность темы исследования. Во многих прикладных задачах возникает необходимость обработки больших наборов символьных последовательностей переменной длины методами машинного обучения. Например, многие вопросы прикладной биоинформатики сводятся к задачам классификации, регрессии и кластеризации на множествах строк. Решению подобных проблем посвящено множество исследований. В качестве примера можно привести работу Х. Сайго и Д. Верта, Н. Уэда и Т. Акутсу 2004 года[1], в которой метод опорных векторов в сочетании со строковым ядром на основе локального выравнивания используется для поиска гомологичных участков белковых последовательностей. Также в последнее время большое применение в биоинформатике получили алгоритмы на основе глубокого обучения. Подробный обзор таких исследований дан М. Вайнбергом, Д. Мерико, А. Делонгом и Б. Д. Фрейем в работе, опубликованной в 2018 году[2] в журнале Nature Biotechnology.

В отличие от задач обработки многомерных векторов и растровых изображений применение методов машинного обучения на строковых данных сопряжено с рядом сложностей. Это можно проиллюстрировать на примере изучаемой в диссертации задачи поиска строковой медианы, которая представляет собой следующую проблему дискретной оптимизации:

$$m(D) = \arg\min_{a \in G^*} \sum_{b \in D} \mathrm{ED}(a,b), \tag{1}$$

где G — алфавит, G^* — множество строк над алфавитом G произвольной конечной длины, $D \subset G^*$ — конечный набор строк, $\mathrm{ED}(a,b)$ — редакционное расстояние Левенштейна, которое определяется как минимальное количество вставок, замен и удалений одного символа, необходимых чтобы преобразовать строку a в строку b.

Впервые функция ED была предложена в 1965 году В. И. Левенштейном для случая бинарного алфавита. Впоследствии в 1974 году Р. Вагнер и М. Фишер обобщили метрику Левенштейна на случай произвольного алфавита, а также предложили полиномиальный алгоритм динамического программирования для ее вычисления. Редакционное расстояние Левенштейна относится к широкому классу методов выравнивания последовательностей, который также включает и другие строковые метрики, используемые для сравнения последовательностей.

[1] Protein homology detection using string alignment kernels / H. Saigo [и др.] // Bioinformatics. 2004. Т. 20, № 11. С. 1682—1689.

[2] Deep learning in biomedicine / M. Wainberg [и др.] // Nature biotechnology. 2018. Т. 36, № 9. С. 829.

Задача поиска строковой медианы (1) играет ключевую роль во многих приложениях биоинформатики например при кластеризации последовательностей ДНК. Кроме того, к проблемам оптимизации, подобным (1), сводятся многие задачи машинного обучения на строковых данных, например задача классификации строк на основе коротких подпоследовательностей, называемых в биоинформатике мотивами.

В работе Ц. Хигуэра 2000 года[3] было доказано, что проблема (1) является NP-сложной. В 1997 году Ф. Касакуберта и М. Антонио предложили жадный алгоритм, позволяющий находить приближенное решение задачи поиска строковой медианы за полиномиальное время. Предложенный ими метод дает хорошие результаты во многих приложениях, однако является недостаточно гибким. В частности, он не позволяет решать задачи с ограничениями на строки. Кроме того, подход, лежащий в основе жадного алгоритма, не является универсальным и не подходит для модификаций задачи (1), в которых используются метрики на основе локального выравнивания, например его нельзя применить в упомянутой выше задаче классификации строк на основе мотивов.

Целью данной работы является разработка моделей гладких строковых (квази)метрик, которые позволяют решать задачу поиска строковой медианы и подобные ей задачи оптимизации с помощью градиентных методов оптимизации, а также создание на основе этих моделей эффективных алгоритмов машинного обучения на строковых данных. Поскольку данные метрики основаны на выравнивании последовательностей, предложенные в диссертации модели и численные методы собирательно называются методами дифференцируемого выравнивания.

Решены следующие задачи:

1. В задаче поиска медианы в пространстве (позиционно) вероятностных матриц (ВМ), представляющих строки, доказана теорема о достижимости минимума на бинарной ВМ.
2. На основе теоремы достижимости разработана модель гладкой аппроксимации расстояния Левенштейна, называемая дифференцируемым редакционным расстоянием (ДРР), и найдены рекуррентные формулы для эффективного вычисления ее значений и градиента.
3. Предложен эффективный численный метод для приближенного поиска строковой медианы.
4. Разработана модель дифференцируемой меры схожести мотива (короткой подпоследовательности) и последовательности (ДМС), а также модель классификатора последовательностей на основе этой меры.

[3] *Higuera C. de la.* Topology of strings: Median string is NP-complete / C. de la Higuera, F. Casacuberta // Theoretical computer science. 2000. Т. 230, № 1/2. С. 39—48.

5. Разработаны методы визуализации мотивов, найденных в процессе обучения классификатора.
6. Разработано обобщение алгоритма кластеризации K-средних для строковых данных с поиском медианы при помощи ДРР. Предложенный алгоритм исследован на синтетических данных, а также на примере репертуара Т-клеточных рецепторов.
7. Предложенная модель нейронной сети валидирована на синтетических данных, а также применена задаче предсказания связывания пептидов и молекул MHC второго класса.
8. Разработан комплекс программ, содержащий реализацию предложенных алгоритмов с использованием параллельных вычислений на графическом процессоре.

Научная новизна. Новыми является предложенная модель дифференцируемого редакционного расстояния вместе с рекуррентными формулами для ее расчета и численный метод для поиска строковой медианы при помощи ДРР. Также новыми являются: предложенная модель дифференцируемой меры схожести мотива и последовательности, архитектуры нейронных сетей на основе ДМС, обобщение алгоритма K-средних для строковых данных с поиском медианы при помощи ДРР.

Практическая значимость. Предложенная модель ДРР и численный метод поиска строковой медианы на его основе могут использоваться в различных задачах машинного обучения на строковых данных, например для кластеризации последовательностей методом K-средних. Разработанная модель ДМС, а также архитектура нейронной сети на ее основе могут быть применены в задачах классификации последовательностей на основе мотивов. Программный комплекс состоит из набора подключаемых модулей, которые могут быть использованы в различных прикладных задачах машинного обучения, связанных с обработкой строковых данных.

Методология и методы исследования. В работе используются современные методы машинного обучения, в том числе алгоритмы глубокого обучения. Применяются градиентные методы оптимизации для задач большой размерности. При разработке комплекса программ были задействованы технологии параллельных вычислений на графическом процессоре.

Основные положения, выносимые на защиту:

1. Параметрическая модель дифференцируемого редакционного расстояния (ДРР), аппроксимирующая расстояние Левенштейна и в пределе совпадающая с ним; численный метод расчета ДРР и его градиента на основе выведенных рекуррентных формул.
2. Модель дифференцируемой меры схожести мотива и последовательности (ДМС), учитывающая возможные вставки символов; численный метод для расчета ДМС на основе полученных рекуррентных формул.

3. Эффективный численный метод приближенного поиска строковой медианы при помощи ДРР за полиномиальное время.
4. Численный метод кластеризации символьных последовательностей на базе алгоритма K-средних с поиском медианы при помощи ДРР.
5. Модель слоя нейронной сети для решения задачи классификации строк на основе ДМС и ее обоснование.
6. Программный комплекс, содержащий реализацию предложенных моделей и численных методов, который может использоваться для решения прикладных задач машинного обучения.

Достоверность полученных результатов подтверждается соответствующими математическими доказательствами, вычислительными экспериментами на синтетических и реальных данных, а также согласованностью результатов, полученных разными методами.

Апробация работы. Основные результаты работы докладывались на:

– пятом Европейском конгрессе по иммунологии (Амстердам, Нидерланды, 2–5 сентября 2018);
– международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2019» (Москва, МГУ имени М. В. Ломоносова, 8-12 апреля, 2019);
– XVI международной конференции «Алгебра, теория чисел и дискретная геометрия: современные проблемы, приложения и проблемы истории», посвященной 80-летию со дня рождения профессора Мишеля Деза (Тула, ТГПУ им. Л. Н. Толстого, 13-18 мая, 2019).

Публикации. Основные результаты по теме диссертации изложены в 12 печатных изданиях, 5 из которых изданы в журналах, рекомендованных ВАК, 7 — в тезисах докладов всероссийских и международных конференций. Получено свидетельство о государственной регистрации программы для ЭВМ № 2019615354.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы.

В первой главе диссертации предлагается новый алгоритм для приближенного решения задачи поиска строковой медианы с использованием ДРР. Смысл данного метода заключается во вложении дискретной задачи (1) в эквивалентную ей гладкую задачу. При этом приближенное решение гладкой задачи может быть эффективно найдено с помощью градиентных

методов оптимизации (стохастического градиентного спуска с проекцией градиента). Данный подход строго обосновывается, в частности, в задаче о медиане доказывается совпадение решений в дискретной и гладкой задачах.

Перейдем к деталям реализации. Пусть $G = \{g_1, \ldots, g_n\}$ — конечное множество различных символов — алфавит, $n = |G|$, G^* — пространство всевозможных последовательностей символов или строк $x = x(1)x(2)\ldots x(L)$ над G произвольной длины $L = |x| \geq 0$. Естественной метрикой в G^* является редакционное расстояние Левенштейна $\mathrm{ED}(x_1,x_2)$. Оно определяется как минимальное количество вставок, замен и удалений символов, необходимых для преобразования одной строки в другую.

Основная идея предложенного метода поиска медианы состоит в представлении последовательностей $x(1)x(2)\ldots x(L) \in G^*$ вероятностными матрицами (ВМ) $X = (X(i,j)) \in [0,1]^{L\times n}$, строки которых $X(i)$ являются вероятностными (или стохастическими) векторами, т. е. $\sum_{j=1}^{n} X(i,j) = 1$, $i = \overline{1,L}$, $L \geq 0$ ($L = 0$ отвечает пустой матрице $\emptyset$). Отметим, что в данном случае термин «вероятностный» взят из работы[4] и означает формальное соответствие строк матрицы требованию $\sum_{j=1}^{n} X(i,j) = 1$, $X(i,j) \in [0,1]$.

Множество таких матриц для произвольного $L \geq 0$ обозначим через V_n. Вначале строки представляются бинарными ВМ, содержащими только нули или единицы, множество которых обозначается $V_{n,b}$. Каждой строке $x = x(1)\ldots x(L) \in G^*$ биективно соответствует бинарная ВМ $X \in V_{n,b}$, где $X(i,j) = 1$, если $x(i) = g_j$, и $X(i,j) = 0$ иначе. Поэтому в дальнейшем не делается различий между строками и их представлениями в виде бинарных ВМ. Также как и в случае последовательностей используются строковые обозначениями $X = X(1)\ldots X(L)$, где $L = |X|$ — называется длиной ВМ, $X(i)$ — вектор строки матрицы X.

Используя данную параметризацию, расстояние Левенштейна, можно определить как функцию, заданную на $V_n \times V_n$. Для этого вводится понятие соответствия двух ВМ

Определение 1. *Соответствием двух ВМ X_1,X_2 называется произвольная пара (X_1',X_2'') подматриц равной длины $|X_1'| = |X_2''| = l$:*

$$X_1' = X_1(i_1')\ldots X_1(i_l') \subseteq X_1, \quad X_2'' = X_2(i_1'')\ldots X_2(i_l'') \subseteq X_2.$$

Здесь l называется длиной соответствия.

Справедливо следующее утверждение, позволяющее определить редакционное расстояние Левенштейна в терминах соответствий.

---

[4] *Leung H. C.* Finding exact optimal motifs in matrix representation by partitioning / H. C. Leung, F. Y. Chin // Bioinformatics. 2005. Т. 21, suppl_2. С. ii86—ii92.

Утверждение 1. *Пусть* $X_1, X_2 \in V_{n,b}$ *— матричные представления строк* $x_1, x_2 \in G^*$, $L_1 = |X_1|$, $L_2 = |X_2|$. *Тогда*

$$\mathrm{ED}(X_1, X_2) = \min \left\{ \frac{1}{2} \|X_1' - X_2''\|_1 + (L_1 - l) + (L_2 - l) \right\}. \tag{2}$$

Здесь минимум берется по всем (матричным) соответствиям (X_1', X_2''), $l = |X_1'| = |X_2''|$, *т. е. подматрицам вида* $X_1' = X_1(i_1') \dots X_1(i_l')$, $X_2'' = X_2(i_1'') \dots X_2(i_l'')$,

$$\|X_1' - X_2''\|_1 = \sum_{a=1}^{l} \|X_1'(a) - X_2''(a)\|_1, \tag{3}$$

где $X_1'(a) = X_1(i_a')$, $X_2''(a) = X_2(i_a'')$.

Легко видеть, что в случае, если X_1 и X_2 — бинарные ВМ, то определенная выше функция $\mathrm{ED}(X_1, X_2)$ в точности совпадает с расстоянием Левенштейна между строками x_1 и x_2. Этот факт позволяет сформулировать задачу о поиске строковой медианы как проблему оптимизации на $V_{n,b}$.

Определение 2. *Задача поиска строковой медианы* $M(D)$ *для конечного набора строк* D *по метрике* ED *может быть сформулирована в следующем матричном виде:*

$$S(D,U) = \min_{X \in U} \sum_{k=1}^{K} \mathrm{ED}(D_k, X), \quad M(D) = \arg S(D,U), \tag{4}$$

где $U = V_{n,b}$ *и* $D = \{D_k\}_{k=1}^{K} \subset V_{n,b}$ *— набор бинарных ВМ, представляющих строки.*

Заметим, что утверждение 1 позволяет естественным образом определить функцию $\mathrm{ED}(X_1, X_2)$ для произвольных вещественных позиционно вероятностных матриц $X_1, X_2 \in V_n$.

Нам потребуется операция округления $\mathrm{round}(X)\colon V_n \to V_{n,b}$. Оно определяется построчно: $\mathrm{round}(X) = \mathrm{round}(X_1) \dots \mathrm{round}(X(L))$. Округлением вектор-строки $X(i)$ называется бинарная строка длины n, в которой максимальный из элементов $X(i,j)$ заменяется единицей (если таких несколько, то, например, первый), а остальные $X(i,j)$ — нулями. Несложно показать, что округление отвечает наилучшему приближению вектора $X(i)$ по ℓ_1-метрике вероятностным бинарным вектором.

Минимум в задаче (2) достигается на некоторой бинарной ВМ $B \in V_{n,b}$. Следующая теорема показывает, что минимум не изменится, если его искать на объемлющем непрерывном пространстве V_b.

Теорема 1. *Имеем*

$$S(D,V_{n,b}) = S(D,V_n).$$

Кроме того, если $X_* \in V_n$ *— решение задачи* $S(D,V_n)$*, то* $\mathrm{round}(X_*) \in V_{n,b}$ *также решение* $S(D,V_n)$*, а, значит, и решение* $S(D,V_{n,b})$*.*

Доказательство данной теоремы, приведено в диссертации. Она позволяет решать дискретную задачу оптимизации (4) на множестве произвольных вероятностных матриц. Однако функция $\mathrm{ED}(X_1,X_2)$ по прежнему определяется через не гладкую операцию минимума. Чтобы использовать при решении задачи (4) эффективные численные методы оптимизации первого порядка в диссертации предлагается модель гладкой аппроксимации расстояния Левенштейна, называемая дифференцируемым редакционным расстоянием (ДРР).

Для построения ДРР предлагается заменить минимум в выражении (2) на экспоненциально-взвешенную сумму, в машинном обучении называемую $\mathrm{soft\,min}$. Для конечного множества вещественных чисел D операция $\mathrm{soft\,min}$ определяется как

$$\mathrm{soft\,min}(D;\tau) = \frac{\sum_{x\in D} xe^{\tau x}}{\sum_{x\in D} e^{\tau x}},$$

где $\tau < 0$ — параметр аппроксимации (в дальнейшем для краткости может опускаться). Известно, что $\mathrm{soft\,min}(D;\tau)$ стремится к $\min D$ экспоненциально быстро при $\tau \to -\infty$ (см. диссертацию).

Это приводит к следующему определению:

Определение 3. *Дифференцируемое редакционное расстояние между двумя последовательностями в матричном представлении* $X_1,X_2 \in V_n$ *с длинами* L_1*,* L_2 *соответственно определяется следующим выражением:*

$$\mathrm{SED}(X_1,X_2) = \mathrm{soft\,min}\left\{\frac{1}{2}\|X_1' - X_2'\|_1 + (L_1 - l) + (L_2 - l)\right\}, \qquad (5)$$

где $\mathrm{soft\,min}$ *берется по всем соответствиям* (X_1',X_2'')*,* $l = |X_1'| = |X_2''|$*.*

При $\tau \to -\infty$ ДРР стремится к расстоянию Левенштейна, что следует из свойств функции $\mathrm{soft\,min}$. В общем случае скорость сходимости зависит от длин сравниваемых последовательностей. Экспериментально было проверено, что для строк длиной до 30 символов уже при $\tau = -3$ коэффициент детерминации R^2 между истинным и дифференцируемым редакционными расстояниями превосходит 0.99. (см. таблицу 1 в диссертации).

Раздел 1.2.3 посвящен эффективному вычислению метрики SED. На практике использовать для расчета SED выражение (5) невозможно, так как это требуется перебора соответствий, общее число которых растет экспоненциально с ростом длин строк X_1 и X_2.

В работе предлагается подход, позволяющий вычислить SED за полиномиальное время при помощи рекуррентных формул. Для этого вводятся следующие вспомогательные матрицы:

$$\begin{aligned}\alpha(i,j) &= \sum_{(X_1',X_2'')\in\Omega_{i,j}} R_{i,j}(X_1',X_2'')e^{\tau R_{i,j}(X_1',X_2'')},\\ \beta(i,j) &= \sum_{(X_1',X_2'')\in\Omega_{i,j}} e^{\tau R_{i,j}(X_1',X_2'')}, \quad i=\overline{0,L_1},\ j=\overline{0,L_2}.\end{aligned} \tag{6}$$

Здесь $\Omega_{i,j}$ — множество всех матричных соответствий (X_1', X_2'') подматриц $X_1(1:i)$ и $X_2(1:j)$,

$$R_{i,j}(X_1',X_2'') = \frac{1}{2}\|X_1'-X_2''\|_1 + i - l + j - l, \quad l = |X_1'| = |X_2''|.$$

Тогда

$$\mathrm{SED}(X_1,X_2) = \frac{\alpha(L_1,L_2)}{\beta(L_1,L_2)}. \tag{7}$$

Теорема 2. *Верны следующие рекуррентные формулы: для* $\tau<0$ $i=\overline{1,L_1}$, $j=\overline{1,L_2}$

$$\begin{aligned}\alpha(i,j) = \big\{\alpha(i-1,j)+\beta(i-1,j)+\alpha(i,j-1)+\beta(i,j-1)\big\}e^{\tau} +\\ +\big\{\alpha(i-1,j-1)+\beta(i-1,j-1)\delta(i,j)\big\}e^{\tau\delta(i,j)} -\\ -\big\{\alpha(i-1,j-1)+2\beta(i-1,j-1)\big\}e^{2\tau},\end{aligned}$$

$$\beta(i,j) = \big\{\beta(i-1,j)+\beta(i,j-1)\big\}e^{\tau} + \beta(i-1,j-1)\big\{e^{\tau\delta(i,j)}-e^{2\tau}\big\},$$

$$\delta(i,j) = \frac{1}{2}\|X_1(i)-X_2(j)\|_1,$$

где для $i=\overline{0,L_1}$, $j=\overline{0,L_2}$

$$\alpha(i,0)=ie^{\tau i}, \quad \alpha(0,j)=je^{\tau j}, \quad \beta(i,0)=e^{\tau i}, \quad \beta(0,j)=e^{\tau j}.$$

Рекуррентные формулы из теоремы 2 дают эффективный способ вычисления ДРР за полиномиальное время. Дифференцирование этих формул позволяет получить рекуррентные соотношения для расчета частных производных $\frac{\partial\alpha(i,j)}{\partial X_k(i,j)}$, $\frac{\partial\beta(i,j)}{\partial X_k(i,j)}$, $k=1,2$, что дает возможность эффективно вычислять градиент ДРР $\nabla\,\mathrm{SED}(X_1,X_2)$ по X_1 или X_2 (см. детали в диссертации). В итоге это позволяет приближенно решать задачу (2) за полиномиальное время.

Вторая глава диссертации посвящена решению задачи классификации (регрессии) на строковых данных на основе коротких непустых подпоследовательностей, называемых в биоинформатике мотивами. В формальной постановке эта задача заключается в построении модели классификатора

Таблица 1 — Пример обучающей выборки в задаче классификации строк на основе коротких мотивов

$y = 0$	$y = 1$
GCTGAC	TACTCA
TAGACC	AGGCAT

$M_{\mathbf{w},\mathbf{z}}(x)$, выход которого будет зависеть от некоторого набора параметров w и наличия во входной строке $x \in G^*$ одного или нескольких коротких мотивов $z_i \in G^*$, $|z_i| < |x|$, $i = \overline{1,m}$. В таблице 1 приведен простой пример, иллюстрирующий идею строковой классификации при помощи мотивов.

Таблица 1 содержит пример обучающей выборки D_{train}. В первом и втором столбцах находятся строки относящие к классу 0 и классу 1 соответственно. Значение y зависит от наличия в последовательности одного из двух мотивов: $z_1 = \text{TGA}$ для $y = 0$ и $z_2 = \text{ACT}$ для $y = 1$. Основной сложностью является то, что эти мотивы заранее не известны и определяются в процессе обучения. Напомним, что мотивы являются подпоследовательностями, и, значит, могут содержать разрывы. Поэтому при определении присутствия мотива в строке требуется учитывать вставки символов.

В общем виде модель строкового классификатора на основе мотивов $M_{W,\mathbf{z}}$ представляется в виде

$$M_{\mathbf{w},\mathbf{z}}(x) = \text{Cl}_W\left(F_{z_1}(x), \ldots, F_{z_m}(x)\right), \quad x \in G^*. \tag{8}$$

Здесь $F_z(x)\colon G^* \to \mathbb{R}$ — мера сходства входной последовательности x и мотива $z \in G^*$, формирующая вектор признаков классификации и показывающая, содержит ли строка x подпоследовательность z. $\mathbf{z} = (z_1, \ldots, z_m) \in (G^*)^m$ — вектор мотивов. Функция $\text{Cl}_{\mathbf{w}}\colon \mathbb{R}^m \to Y$ — финальный классификатор $\text{Cl}_{\mathbf{w}}$, где $\mathbf{w}$ — набор дополнительных параметров, определяемых архитектурой $\text{Cl}_{\mathbf{w}}$.

Параметры $\mathbf{w}$ и $\mathbf{z}$ определяются в процессе обучения модели. Оно заключается в минимизации на обучающей выборке $D_{\text{train}} = \{(x_1,y_1), \ldots, (x_N,y_N)\} \subset G^* \times Y$ ошибки модели, определяемой выражением

$$\frac{1}{N} \sum_{(x,y) \in D_{\text{train}}} \text{Loss}\left(M_{\mathbf{w},\mathbf{z}}(x), y\right) \to \min_{\mathbf{w},\mathbf{z}}.$$

Здесь $\text{Loss}\colon Y \times Y \to \mathbb{R}_+$ — так называемая лосс-функция, показывающая на сколько близко предсказанное значение $M_{\mathbf{w},\mathbf{z}}(x)$ к истинному y.

Можно рассмотреть произвольную архитектуру классификатора $\text{Cl}_{\mathbf{w}}$: линейный классификатор, нейронная сеть, решающее дерево и т. п. Общим моментом в них является необходимость построения «хорошей» функции F_z,

с одной стороны, обеспечивающей высокую точность классификатора, а с другой — дающей возможность его эффективного обучения.

Подобный подход к классификации последовательностей является востребованным в биоинформатике, поскольку классификатор вида (8) дает возможность явно выделить (визуализировать) найденные в процессе обучения мотивы z и их соответствие меткам y.

Во второй главе мы предлагаем модель строкового классификатора, в котором в качестве $\mathrm{Cl}_{\mathbf{w}}$ используется простой линейный классификатор (логистическая регрессия), а функция F_z строится на основе модели дифференцируемой меры схожести мотива и последовательности (ДМС).

Для ее построения мы также, как и в первой главе представляем последовательности в виде ВМ. Для параметризации неизвестного мотива длины K, используется так называемая позиционно-весовая матрица $\Theta = \Theta(1),\ldots,\Theta(K)$, $\Theta = (\Theta(i,j)) \in \mathbb{R}^{K\times|G|}$, в которой $\Theta(i,j)$ оценивает вероятность того, что i-й символ мотива является j-м символом алфавита. ПВМ мотива связана с его вероятностной матрицей Z операцией $\mathrm{soft\,arg\,max}$ (см. диссертацию). Также для параметризаци используется вектор штрафов $g = (g(1),\ldots,g(K-1)) \in \mathbb{R}_{-}^{K-1}$. Элемент $g(i)$ в данном случае равняются штрафу за вставку символа на i-й позиции мотива, а $e^{g(i)}$ имеет смысл вероятности соответствующей вставки.

Для построения ДМС мы применяем подход на основе выравниваний, аналогичный тому, что ранее использовался при определении ДРР. Получаемая метрика $\mathrm{SF}_{\Theta,g}$ позволяет учитывать вставки символов и может быть использована для определения функции F_z в формуле (8). При этом, благодаря свойству дифференцируемости по элементам ПВМ Θ и вектора штрафов g, классификатор на ее основе может обучаться с помощью стандартных градиентных методов.

Также как и в случае ДРР, для расчета значений и градиента ДМС используются полиномиальный алгоритм на основе рекуррентных формул.

Мы используем ДМС для определения модели линейного бинарного классификатора:

$$M_{\mathbf{w},\boldsymbol{\Theta},\mathbf{g}}(X) = \sigma\left(w_0 + \sum_{i=1}^{m} w_i A\left(\mathrm{SF}_{\Theta_i,g_i}(X)\right)\right), \tag{9}$$

где $\mathrm{SF}_{\Theta_i,g_i}\colon V_{|G|} \to \mathbb{R}$ — предложенная дифференцируемая мера схожести между мотивом с параметром Θ_i и строкой с ВМ X, $A\colon \mathbb{R}\to\mathbb{R}$ — некоторая стандартная активационная функция, w_i — весовые коэффициенты финального слоя, $\sigma(t) = \frac{1}{1+e^{-t}}$ — логистическая функция.

В целом описанный в диссертации подход является универсальным и может быть применен и в других подобных задачах машинного обучения на строковых данных. Например, предложенная модель ДМС может использоваться для построения нейронных сетей.

Также во второй главе в разделе 2.5 предлагается два алгоритма визуализации, которые позволяют интерпретировать полученные в результате обучения ПВМ мотивов. Первый метод основан на визуализации вероятностной матрицы мотива Z. В другом используются так называемые карты выравнивания, которые выделяют в исходной строке символы последовательности, соответствующие мотиву.

Третья глава посвящена применению моделей, описанных в первых двух главах в задачах машинного обучения на строковых данных. Первой такой задачей является задача кластеризации строк с помощью метода K-средних (см. диссертацию раздел 3.1). Предложенный метод поиска строковой медианы при помощи ДРР позволяет явным образом применить алгоритм K-средних к строковым данным. В отличие от других методов кластеризации последовательностей, использующих матрицу расстояний, разработанный алгоритм позволяет явно находить медианы кластеров, что является важным во многих приложениях. Для тестирования предложенного подхода проводится ряд экспериментов на синтетических и реальных данных, результаты которых приводятся в диссертации в таблице 5.

Также в данной главе подробно изучается возможность использования моделей на основе ДМС в задачах классификации и регрессии на строковых данных. В диссертации проводится ряд экспериментов на синтетических данных, подтверждающих, что предложенная во второй главе модель классификатора может правильно находить мотивы, разделяющие различные классы последовательностей (см. таблицу 7).

Кроме этого, в данной главе нейронная сеть на основе ДМС применяется в реальной задаче предсказания связывания коротких пептидов и молекул MHC второго класса. Эта задача играет важную роль в изучении работы иммунной системы, а также при разработке новых вакцин. С математической точки зрения данная проблема формулируется, как задача регрессии на множестве строк над алфавитом из 20 букв. Для ее решения была использована нейронная сеть на основе ДМС. В результате такой подход показал точность близкую к лучшей на сегодняшний день модели — NetMHCpanII[5]. При этом в отличии от NetMHCpanII предложенная модель хорошо интерпретируема, за счет использования архитектуры на основе ДМС. Применяя описанный во второй главе алгоритм визуализации нам удалось выделить специфические мотивы, ответственные за связывание пептидов с определенными комплексами MHC.

Четвертая глава диссертации посвящена программной реализации моделей и численных методов, описанных в предыдущих главах. Основу разработанного программного комплекса составляет реализация алгоритмов вычисления значений и градиента ДРР и ДМС. При создании этой

[5]Improved methods for predicting peptide binding affinity to MHC class II molecules / K. K. Jensen [и др.] // Immunology. 2018. Т. 154, № 3. С. 394—406.

реализации мы использовали технологию параллельных вычислений на графическом процессоре CUDA. Приведенные в первых двух главах рекуррентные формулы для дифференцирования ДРР и ДМС могут быть эффективно распараллелены. Также возможность ускорения с помощью параллельных вычислений возникает при расчете значений данных метрик в задачах поиска строковой медианы и обучения классификатора последовательностей за счет пакетной обработки строк. Это позволило нам задействовать ресурсы графического процессора.

Программный код нашей реализации написан на языке Python с использованием библиотеки для глубокого обучения Chainer. Реализация моделей ДРР и ДМС выполнена на основе стандартного интерфейса Chainer, что позволяет легко использовать предложенные модели в сочетании со стандартными инструментами данной библиотеки.

Помимо моделей ДРР и ДМС, разработанный комплекс программ содержит реализацию описанных во второй главе методов визуализации, алгоритма кластеризации строковых данных и моделей классификаторов, приведенных в третьей главе. Данные реализации имеют единообразный функциональный интерфейс и могут быть использованы в пользовательских скриптах.

Исходный код реализации доступен в GitHub[6].

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

- Предложенный численный метод поиска строковой медианы при помощи ДРР превосходит по точности существующие подходы. Кроме того, данный подход является универсальным и может быть применен для различных модификаций задачи о поиске строковой медианы, а также в других подобных проблемах оптимизации. Это дает возможность применять многие классические алгоритмы машинного обучения, например метод K-средних к строковым данным.
- Разработанная в работе мера сходства мотива и последовательности ДМС, а также модель классификатора на её основе позволяет эффективно решать задачу классификации строк на основе мотивов.
- Разработанная в работе модель слоя поиска мотивов может быть использована для построения интерпретируемых нейронных сетей. В задаче предсказания связывания пептидов и молекул MHC второго класса, такая сеть показала точность, близкую к лучшей на текущей момент модели NetMHCpanII. При этом в отличии от NetMHCpanII, предложенная архитектура позволяет визуализировать мотивы связывания.

[6] *Ofitserov E.* Soft edit distance / E. Ofitserov. 2019. URL: https://github.com/JenEskimos/soft_edit_distance (дата обр. 03.06.2019).

– Созданный комплекс программ интегрирован с фреймворком для глубокого обучения Chainer и позволяет использовать предложенные в работе модели и методы машинного обучения в комбинации со стандартным функционалом этой библиотеки для написания пользовательских скриптов. Исходный код с открытой лицензией доступен на GitHub.

В перспективе планируются продолжить исследования по следующим направлениям:

– Расширить модель ДРР для случая переменной цены замены символа, а также разработать дифференцируемую версию алгоритма локального выравнивания Смита–Ватермана.
– Оптимизировать алгоритмы расчета ДРР и поиска мотивов для работы с длинными последовательностями.
– Применить предложенные в работе методы поиска строковой медианы и кластеризации в задачах связанных со сборкой генома.

Публикации автора по теме диссертации

1. *Офицеров Е. П.* Статистическая модель периферической селекции Т-клеточных рецепторов / Е. П. Офицеров // Известия ТулГУ. Тех. науки. — 2017. — № 2. — С. 138—143.
2. *Офицеров Е. П.* Глубокая модель селекции Т-клеточных рецепторов / Е. П. Офицеров // Известия ТулГУ. Тех. науки. — 2017. — № 12—2. — С. 350—355.
3. *Офицеров Е. П.* Классификация последовательностей на основе коротких мотивов / Е. П. Офицеров // Чебышевский сборник. — 2018. — Т. 19, № 1. — С. 187—199. — DOI: 10.22405/2226-8383-2018-19-1.
4. *Офицеров Е. П.* Программный комплекс для решения задач машинного обучения на строковых данных при помощи дифференцируемого редакционного расстояния / Е. П. Офицеров // Известия ТулГУ. Тех. науки. — 2019. — № 5. — С. 370—376.
5. *Горбачев Д. В.* Новый подход к поиску строковой медианы и визуализация строковых кластеров / Д. В. Горбачев, Е. П. Офицеров // Чебышевский сборник. — 2019. — Т. 20, № 2. — С. 85—99. — DOI: 10.22405/2226-8383-2019-20-2-85-99.
6. *Офицеров Е.* Статистический вывод параметров селекции иммунных рецепторов с помощью алгоритма градиентного спуска с переменным шагом / Е. Офицеров, А. Воеводская, В. Назаров // Новые информационные технологии в автоматизированных системах. — 2016. — № 19. — С. 149—155.

7. Deep learning model for prediction of probability for thymic selection for T-cell receptor sequences / S. Tolstoukhova [et al.] // Proceedings of Moscow Conference on Computational Molecular Biology. — 2017.

8. *Tolstoukhova S.* Deep learning model of clonal selection for T-cell receptor sequences / S. Tolstoukhova, E. Ofitserov, V. Nazarov // Abstracts of European Conference on Computational Biology. — 2017.

9. *Nazarov V.* Visualisation of immune receptor sequences via projection into high-dimensional space with Siamese neural networks / V. Nazarov, E. Ofitserov, V. Tsvetkov // Abstracts of the 5th European Congress of Immunology — ECI 2018 - Amsterdam, The Netherlands. — 2018. — P. 111.

10. *Ofitserov E.* TCR sequence motif based classification of CD4-CD8 cells / E. Ofitserov, V. Tsvetkov, V. Nazarov // Abstracts of the 5th European Congress of Immunology — ECI 2018 - Amsterdam, The Netherlands. — 2018. — P. 520.

11. *Nazarov V.* Robust prediction of peptide-MHC binding affinity with deep neural networks / V. Nazarov, V. Tsvetkov, E. Ofitserov // Abstracts of the 5th European Congress of Immunology — ECI 2018 - Amsterdam, The Netherlands. — 2018. — P. 517.

12. *Офицеров Е. П.* Кластеризация последовательностей с помощью дифференцируемого редакционного расстояния / Е. П. Офицеров // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2019» секция «Биоинженерия и Биоинформатика» подсекция «Биоинформатика» [Электронный ресурс]. — 2019.

13. *Ofitserov E.* Soft edit distance for differentiable comparison of symbolic sequences / E. Ofitserov, V. Tsvetkov, V. Nazarov // arXiv preprint arXiv:1904.12562. — 2019.

14. *Офицеров Е. П.* Программный комплекс для обработки строк методами дифференцируемого выравнивания // Сведетельство о государственной регистрации программы для ЭВМ 2019615354 / Е. П. Офицеров. — 2019.