DOI: 10.17323/2587-814X.2020.1.19.31

Анализ клинических путей пациентов в медицинских учреждениях на основе методов жесткой и нечеткой кластеризации

Е.С. Прокофьева[a]
E-mail: prokofyeva.liza@gmail.com

Р.Д. Зайцев[b]
E-mail: Roman.Zaitsev@fors.ru

[a] Национальный исследовательский университет «Высшая школа экономики»
Адрес: 101000, г. Москва, ул. Мясницкая, д. 20

[b] ГК «ФОРС»
Адрес: 129272, Москва, ул. Трифоновский тупик, д. 3

Аннотация

Моделирование процессов в системе здравоохранения играет большую роль для понимания ее деятельности и служит основой для повышения эффективности работы медицинских учреждений. Задачи анализа и моделирования больших массивов данных городского здравоохранения с помощью методов машинного обучения представляют особую значимость и актуальность для развития отраслевых решений в рамках цифровизации экономики, где данные являются ключевым фактором производства. В статье рассматривается проблема автоматического анализа и определения групп клинических путей пациентов на основе методов кластеризации. Существующие в данной области работы отражают большой интерес со стороны научного сообщества к подобным исследованиям, однако имеется необходимость развития ряда методологических подходов для дальнейшего их практического применения в городских поликлинических учреждениях с учетом специфики их организации. Целью исследования является повышение качества управления и сегментации входного потока пациентов в городских медицинских учреждениях на основе методов кластерного анализа для дальнейшей разработки рекомендательных сервисов. Одним из подходов для достижения поставленной цели является разработка и внедрение клинических путей, или траекторий движения пациентов. В общем виде под клиническим путем пациента понимается траектория его движения при получении медицинской услуги в соответствующих учреждениях. Представлен подход формирования групп маршрутов пациентов при помощи иерхического агломеративного алгоритма с методом связи Уорда и аддитивной регуляризацией тематических моделей (ARTM). Проведен вычислительный эксперимент на основе данных о маршрутах пациентов с диагнозом сепсис, размещенных в открытом доступе. Особенностью предлагаемого подхода является не только автоматизация определения схожих групп траекторий пациентов, но и учет шаблонов клинических путей для формирования рекомендаций в области организации структуры медицинского учреждения. В результате сформированный подход сегментации входного гетерогенного потока пациентов в городских медицинских учреждениях на основе кластеризации состоит из следующих шагов: 1) подготовка данных медицинского учреждения в формате журнала событий; 2) кодирование маршрутов пациентов; 3) определение верхнего предела длины рассматриваемого пути; 4) иерархическая агломеративная кластеризация; 5) аддитивная регуляризация тематических моделей (ARTM); 6) выявление популярных шаблонов

маршрутов пациентов. Полученные кластеры маршрутов служат основой для дальнейшей разработки имитационной модели медицинского учреждения и предоставления рекомендаций пациентам. Кроме того, эти группы могут быть положены в основу разработки системы «Robotic process automation» (RPA), симулирующей действия человека и позволяющей автоматизировать интерпретацию данных для управления ресурсами учреждения.

Ключевые слова: кластерный анализ; данные; иерархическая кластеризация; тематическое моделирование; коэффициент силуэта; здравоохранение; клинические пути; интеллектуальный анализ процессов.

Цитирование: Прокофьева Е.С., Зайцев Р.Д. Анализ клинических путей пациентов в медицинских учреждениях на основе методов жесткой и нечеткой кластеризации // Бизнес-информатика. 2020. Т. 14. № 1. С. 19–31.
DOI: 10.17323/2587-814X.2020.1.19.31

Введение

Стремительно развивающиеся технологии анализа данных играют огромную роль в здравоохранении. Существующий уровень автоматизации медицинского обслуживания позволяет обрабатывать большие массивы информации и использовать накопленные данные для решения оптимизационных задач. Медицинские учреждения располагают данными о приемах, однако традиционный подход документирования посещений не позволяет сформировать полную картину основных траекторий пациентов и проводить их автоматический анализ.

Важной областью технологий обработки данных в здравоохранении является оптимизация работы медицинских учреждений: эффективный график работы медицинского персонала, прогнозирование потока пациентов, планирование и распределение ресурсов, сокращение очередей и т.д. Современные аналитические технологии позволяют разрабатывать инструменты для принятия решений, в основе которых лежат эмпирические данные. Например, агрегированные данные фактических перемещений пациентов между медицинскими учреждениями и специалистами в рамках этих учреждений позволяют планировать загруженность ресурсов, обеспечивать высокий уровень доступности услуг и оптимизировать работу организации, исходя из реального спроса на эти услуги. На основе таких данных активно развиваются информационные системы управления потоками пациентов [1, 2].

Разработка и внедрение клинических путей (clinical pathways), или траекторий движения пациентов, является важным инструментом в управлении здравоохранением. В общем виде под клиническим путем пациента понимается траектория его движения при получении медицинской услуги в соответствующих учреждениях. Согласно источнику [3], клинические пути стали рассматриваться на международном уровне с 1980-х годов. Данная методология была представлена в медицинских учреждениях Швеции в середине 1990-х. В США, по оценкам источника [3], примерно 80% больниц использовали клинические пути для повышения качества оказываемой помощи.

В исследовании [4] авторы описали клинический путь как план, «отображающий цели для пациентов и определяющий последовательность и время действий, необходимых для достижения этих целей с оптимальной эффективностью».

Согласно источникам [5, 6], разработанные клинические пути интегрировались в электронный документооборот медицинских учреждений. Однако стремительный рост массивов доступных данных, включая изображения в цифровом виде, выявил необходимость автоматического определения клинического пути пациента на основе этих данных. Решением для автоматического выявления персонального клинического пути стали технологии интеллектуального анализа процессов [7, 8], интеллектуального анализа данных [9], алгоритмов машинного обучения [10, 11] и другие.

Существуют различные определения клинического пути. Например, согласно [12], клинический путь может быть определен как структурированный план ухода с указанными основными этапами и сроками лечения пациентов. Важно отметить, что каждая траектория пациента уникальна и соответствует его истории болезни [12]. Клинический путь может включать в себя цепочку соответствующих профилю медицинского учреждения событий: первичную запись на прием к терапевту, лабораторные исследования, получение консультации узкопрофильного специалиста и другие. В работе [13] указано, что выявление шаблонов клинических путей

позволяет потенциально дополнить информацию о намерениях и поведении пациента и также может служить базисом для дальнейшего анализа перемещений пациентов.

В данной статье для решения задачи автоматического анализа и сегментирования клинических путей пациентов рассмотрены методы жесткой и нечеткой кластеризации для повышения стандартизации управления и дальнейшего решения оптимизационных задач развития сервисов для пациентов.

Статья имеет следующую структуру. В разделе 1 приведены существующие подходы к моделированию клинических путей пациента. Раздел 2 содержит формальные определения, общепринятые для моделирования клинических путей, и методологию их кластерного анализа. Вычислительный эксперимент на примере открытых данных больницы и его результаты описаны в разделе 3. В Заключении перечислены тренды и направления дальнейшей работы.

1. Существующие подходы к моделированию клинических путей пациентов

Множество работ [14–18] посвящено исследованию и анализу клинических путей на основе исходных данных медицинского учреждения. Для моделирования клинических путей пациентов может быть применена методологическая база теории вероятностей, математической статистики, интеллектуального анализа данных, теории графов, семантических технологий, интеллектуального анализа процессов и т.п.

В исследовании [14] авторы отметили важность разработки адаптивного подхода при моделировании клинических путей в связи с высокой вариативностью траекторий пациентов и их индивидуальными характеристиками. На основании предложенных графов последовательностей и методов интеллектуального анализа данных (data mining) выделяются шаблоны, или паттерны клинических путей пациентов с инсультом, для прогнозирования траекторий новых пациентов.

Полумарковская модель индивидуальной траектории пациента в клинике семейной практики представлена в работе [15]. Схема общего потока пациентов в этой клинике представлена ориентированным графом, вершинами которого являются кабинеты и отделения клиники, а ребра соответствуют направлению движения анализируемого потока. Данная модель позволяет прогнозировать продолжительность обслуживания пациента, однако такие параметры, как время ожидания в очереди и длина очереди, для анализа недоступны.

Моделирование траекторий пациентов в виде цепей Маркова в работе [16] позволило выявить типичные клинические пути при прогрессировании заболевания и визуализировать их. Благодаря способности Марковских цепей учитывать вложенные модели, в работе [16] клинических путь представлен в виде четырех уровней агрегации. По мнению авторов работы [12], вложенная конструкция позволяет упростить модель клинического пути и выделить наиболее важные закономерности процесса. Однако в качестве основного недостатка цепи Маркова для моделирования клинического пути авторы исследования выделяют ограниченное количество состояний системы.

Впервые метод вероятностного тематического моделирования, в частности модель латентного размещения Дирихле (latent Dirichlet allocation, LDA), был адаптирован для моделирования клинических путей пациентов в исследовании [13], авторы которого предположили, что LDA позволит представить скрытые паттерны лечения пациентов как вероятностные комбинации исходных событий из журнала событий (*рисунок 1*). Латентное распределение Дирихле является порождающей иерархической вероятностной моделью, описанной в 2003 году в исследовании [17] и изначально разработанной для характеристики текстовых документов. Параметры данной модели генерируются из априорного распределения Дирихле, а для обучения модели применяются методы байесовского подхода [17]. Документ в модели LDA представлен распределением по скрытым (латентным) темам, каждая из которых характеризуется распределением по словам. В рамках данного исследования предлагается темы или паттерны полученных тематических моделей называть шаблонами клинических путей.

Особое место среди методов моделирования клинических путей пациентов занимает применение интеллектуального анализа процессов (process mining). Благодаря данному подходу, разработка модели процессов основана на исходных данных о реальном поведении пациентов, их маршрутах и основных характеристиках, влияющих на выбор той или иной траектории. Целью интеллектуального анализа процессов является извлечение новой информации о процессах из журналов событий (логов). Таким образом, интеллектуальный анализ процессов как дисциплина лежит на стыке машинного обучения, ин-

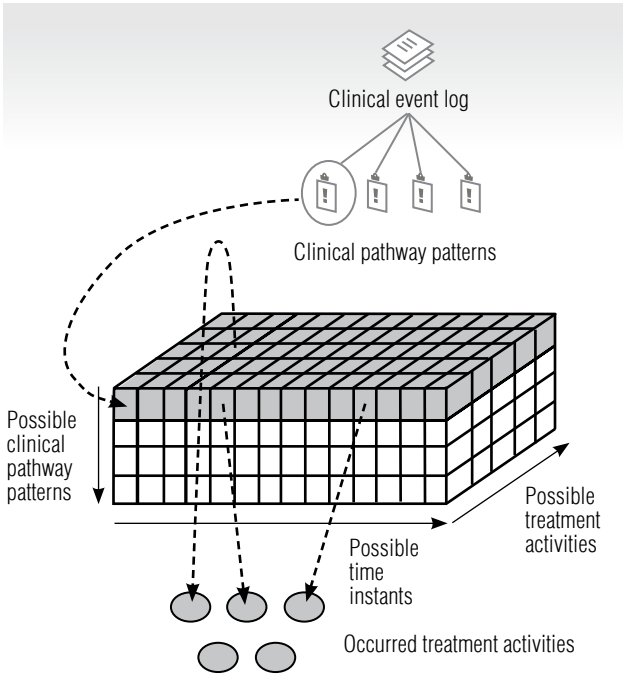

Рис. 1. Исследование шаблонов клинических путей на основе вероятностного тематического моделирования [13]

теллектуального анализа данных и моделирования процессов. Основные положения этой дисциплины изложены в работах [19–22].

Ряд исследований посвящен разработке собственных алгоритмов анализа процессов: например, алгоритма eMotivia для анализа перемещений девяти пациентов в течение 25 недель [23]. Подробный перечень алгоритмов интеллектуального анализа процессов в здравоохранении приведен в работе [7], основанной на обзоре 74 исследований в данной области. Важно отметить, что результаты применения технологий анализа процессов позволяют объективно оценить прошлое и текущее движение потока пациентов, однако для детального изучения поведения системы и проведения экспериментальной части по ее совершенствованию необходима разработка имитационной модели [12, 20, 24].

2. Методы кластерного анализа клинических путей пациентов

2.1. Термины и определения

Характерная структура клинических путей позволяет применять методы кластерного анализа временных рядов. Это обусловлено тем, что ряды, которыми выражены клинические пути, могут иметь как категориальную, так и численную природу. Одним из способов значительно повысить точность разрабатываемой модели является сегментация исходной выборки на подгруппы схожих друг с другом объектов и построения в дальнейшем отдельных «персонифицированных» моделей для каждой из выделенных групп. Исследование направлено на определение таких групп, или кластеров, клинических путей на основе журналов событий медицинских учреждений.

Введем формальные определения, общепринятые для моделирования клинических путей [12–14]. Пусть E – множество всех действительных событий исследуемой области, которые произошли во время процесса медицинского обслуживания: $E \subseteq A \times T$, где A – конечный набор идентификаторов событий, T – множество атрибутов времени. Тогда событие – это пара $e = (a, t)$, где $a \in A$ и $t \in T$. Для обозначения типа активности и отметки времени наступления клинического события используются $e \cdot a$ и $e \cdot t$. Важно отметить, что при моделировании клинических путей каждое событие уникально определено комбинацией его атрибутов. Маршрут σ – событийная цепочка пациента, непустая последовательность событий клинического пути: $\sigma = \langle e_1, e_2, ..., e_n \rangle$, где $e_i \in E$ $(1 \le i \le n)$, $n \in N$ – длина маршрута пациента. Множество всех маршрутов над E обозначается E*. Журнал событий L представляет собой непустое множество маршрутов пациентов над E*: $L = \{\sigma_1, ..., \sigma_m\}$, где $\sigma_i \in E^*$ $(1 \le i \le m)$, $m \in N$.

В рамках данной терминологии теории интеллектуального анализа процессов, адаптированной для моделирования клинических путей пациентов, можно привести следующие примеры соответствия определений:

- маршрут σ_i – пациент, прикрепленный к анализируемой поликлинике;
- событие e_i – посещение терапевта для первичной консультации;
- атрибут a_i – характеристика пациента (пол, возраст, диагноз и т.п.);
- журнал событий L – исходная база данных медицинского учреждения.

Структура журнала событий L предполагает наличие следующих атрибутов [22]:

- идентификатор (patient_id): хранит объекты, для которых выстраиваются последовательности событий;
- деятельность (activity_name): хранит действия, выполняемые в рамках событий журнала;

✦ отметка времени (timestamp): хранит дату и время регистрации событий журнала, например, время посещения терапевта;

✦ ресурс (resource): хранит основных действующих лиц событий журнала (тех, кто выполняет действия в рамках событий журнала). В контексте исследования ресурс может быть представлен медицинским специалистом или оборудованием для проведения исследований;

✦ прочее (other data): остальные данные, которые потенциально полезны для моделирования процессов медицинского учреждения.

2.2. Жесткая кластеризация

Существуют разные системы кодирования маршрутов пациентов. Например, в исследовании [25] авторы используют буквенно-цифровые символы для обозначения активностей клинического пути (могут включать диагнозы, процедуры, анализы и схемы лечения) по стандарту Unicode. Таким образом, авторы указывают на возможность закодировать 65 536 активностей клинического пути.

Выбор системы кодирования зависит от максимального количества активностей журнала событий медицинского учреждения. В данной работе на начальном этапе все маршруты пациентов кодируются путем замены событий на буквы английского алфавита по порядку, поскольку в исходном наборе данных для экспериментальной части представлено 16 видов активностей: ER Registration, ER Triage, IV Liquid, IV Antibiotics, CRP, Admission IC, ER Sepsis Triage, Leucocytes, Lactic Acid, Admission NC, Release A, Release B, Release C, Release D, Release E, Return ER.

В контексте данного исследования, верхний предел длины рассматриваемого пути полагался равным 26 событиям: $Q50 + 3 \cdot (Q75 - Q50)$, где $Q50$ – медиана, а $Q75$ соответствует 75% квантили. Соответственно, аномально длинными считаются пути, содержащие более 26 событий. После анализа распределения длин клинических путей журнала событий аномально длинные пути были исключены из анализа для повышения качества кластеризации.

Для построения кластеризационной модели сравнивались два метода – k-medoids (более устойчивая к выбросам разновидность алгоритма k-means) [26] и иерархический агломеративный алгоритм с типом связи Уорда [27]. Метод Уорда основан на методах дисперсионного анализа для оценки расстояний между группами объектов и, наряду с методом полной связи (complete linkage), приводит к образованию небольших компактных кластеров. Метод применим для задач более дробной классификации объектов с близко расположенными кластерами [29]. Каждый объект выборки в методе Уорда изначально рассматривается как отдельный кластер [30]. На следующем шаге итерации алгоритма объединяются наиболее близкие кластеры, расстояние между которыми измеряется следующей формулой:

$$\Delta(\mathrm{A},\mathrm{B}) = \sum_{i\in \mathrm{A}\cup\mathrm{B}} \left\|\vec{x}_i - \vec{m}_{\mathrm{A}\cup\mathrm{B}}\right\|^2 - \sum_{i\in \mathrm{A}} \left\|\vec{x}_i - \vec{m}_{\mathrm{A}}\right\|^2 - \sum_{i\in \mathrm{B}} \left\|\vec{x}_i - \vec{m}_{\mathrm{B}}\right\|^2 = \frac{n_{\mathrm{A}} n_{\mathrm{B}}}{n_{\mathrm{A}} + n_{\mathrm{B}}} \left\|\vec{m}_{\mathrm{A}} - \vec{m}_{\mathrm{B}}\right\|^2, \quad (1)$$

где A, B – объединяемые кластеры;

$\vec{x}_i$ – объект кластера;

$\vec{m}_{\mathrm{A}\cup\mathrm{B}}$ – центр объединенного кластера AB;

$\vec{m}_{\mathrm{A}}$ – центр кластера A;

$\vec{m}_{\mathrm{B}}$ – центр кластера B;

n_{A} – количество объектов в кластере A;

n_{B} – количество объектов в кластере B.

На основе результатов решения задачи выявления наиболее характерно выделенных кластеров, описанной далее в этом разделе, был выбран алгоритм Уорда.

На следующем этапе была построена матрица расстояний между клиническими путями на основе ограниченного расстояния Дамерау–Левенштейна [31] – меры разницы двух строк символов, определяемой как минимальное количество операций вставки, удаления, замены и перестановки соседних символов, необходимых для перевода одной строки в другую.

Для оценки расстояния Дамерау–Левенштейна между двумя строками a и b определяется функция $d_{a,b}(|a|, |b|)$ [31]:

$$d_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{если } \min(i,j) = 0, \\ \min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \\ d_{a,b}(i-2,j-2)+1 \end{cases} & \text{если } i,j>1, \; a_i = b_{j-1} \text{ и } a_{i-1} = b_j, \\ \min\begin{cases} d_{a,b}(i-1,j)+1 \\ d_{a,b}(i,j-1)+1 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{иначе}, \end{cases} \quad (2)$$

где $1_{(a_i \neq b_j)}$ – индикаторная функция, равная единице при $a_i \neq b_j$ и нулю в противном случае. При этом каждый рекурсивный вызов соответствует одному из следующих случаев:

$d_{a,b}(i-1, j)+1$ – соответствует удалению символа (из a в b),

$d_{a,b}(i, j-1)+1$ – соответствует вставке (из a в b),

$d_{a,b}(i-1, j-1)+1_{(a_i \neq b_j)}$ – соответствие или несоответствие, в зависимости от символов,

$d_{a,b}(i-2, j-2)+1$ – в случае перестановки двух последовательных символов.

Для оценки оптимального количества кластеров применялся коэффициент силуэта (silhouette) [32]. Этот коэффициент основан на идее определения близости каждого исследуемого объекта к своему кластеру. Предположим, что расстояние d на кластеризуемом множестве задано, и с помощью какого-либо метода получена кластеризационная модель. Пусть для каждого объекта выборки i, принадлежащего кластеру C_i, величина $a(i)$ равна среднему расстоянию от i до каждого из объектов j того же кластера:

$$a(i) = \frac{1}{|C_i|} \sum_{i \neq j} d(i, j), j \in C_i. \quad (3)$$

Эта величина косвенно показывает, насколько объект i схож со своим кластером. Далее, назовем кластер C' из множества всех кластеров C соседним для точки i, если

$$C' = \arg\min_{C_k \in C \setminus C_i} \left(\frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \right). \quad (4)$$

Обозначим среднее расстояние от точки i до соседнего кластера как $b(i)$:

$$b(i) = \frac{1}{|C'|} \sum_{j \in C'} d(i, j). \quad (5)$$

Тогда коэффициент силуэта объекта i в полученной модели определяется следующим образом:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}. \quad (6)$$

Видно, что для каждого объекта он изменяется в диапазоне [–1; 1] и показывает, насколько элемент ближе к своему кластеру, чем к ближайшему соседнему. Путем усреднения коэффициентов силуэта элементов можно получить силуэты отдельных кластеров

$$s(C_k) = \frac{1}{|C_k|} \sum_{j \in C_k} s(j) \quad (7)$$

и общий силуэт кластеризационной модели

$$s(C) = \frac{1}{|C|} \sum_{C_k \in C} s(C_k). \quad (8)$$

Наибольший коэффициент силуэта среди кластеризационных моделей, полученных с применением одинакового расстояния d, может быть использован как критерий оптимальности для выбора предпочтительного количества кластеров N и предпочтительного алгоритма кластеризации. Поскольку результаты кластерного анализа должны быть хорошо интерпретируемыми, необходимо выбрать модель, содержащие кластеры с наибольшим силуэтом. Пусть $K_i = \{C_1^i, C_2^i, ...C_{N_i}^i\}$ – кластеризационная модель, разделяющая выборку на N_i кластеров, а $C_{\max}^i$ – кластер с наибольшим силуэтом среди всех кластеров i-й кластеризационной модели:

$$C_{\max}^i = \text{argmax}_{C_j^i \in K_i} \left(s\left(C_j^i\right) \right), j \in 1...N_i. \quad (9)$$

Тогда необходимо найти модель , которая содержит кластер с наибольшим силуэтом среди всех моделей K:

$$K_{opt} = \text{argmax}_{K_i \in K} \left(C_{\max}^i \right). \quad (10)$$

2.3. Нечеткая кластеризация

Вероятностная нечеткая, или перекрывающая, кластеризация маршрутов пациентов по группам клинических паттернов позволяет разработать более гибкий подход при описании общего потока пациентов, где каждый объект выборки относится к кластеру с определенным весом или вероятностью. Применение этого подхода основано на тематическом моделировании, изначально разработанном для определения тематик коллекции текстовых документов. В терминах тематического моделирования события пациента в ходе медицинского обслуживания соотносятся со словами модели. Маршрут пациента представлен последовательностью таких событий аналогично документу со словами. Таким образом, обнаруженные алгоритмом скрытые темы интерпретируются как шаблоны, или паттерны, клинических путей пациента [13].

Согласно исследованию [13], применение метода LDA позволяет для каждого пациента выбрать набор паттернов клинических путей с разными акцентами на значимость этих паттернов. Таким образом, мы моделируем смесь шаблонов маршрутов как полиномиальное распределение вероятности

по клинического пути шаблону z. Аналогично важность каждого клинического действия a при каждом шаблоне моделируется как полиномиальное распределение вероятностей $P(a|\sigma)$ по активностям пациента. Эти два распределения позволяют вычислить вероятность возникновения отдельной клинической активности у пациента:

$$P(a|\sigma) = \sum_{z=1}^{K} P(a|z)\,P(z|\sigma). \qquad (11)$$

В вероятностных порождающих моделях (таких, как, например, LDA) имеющиеся данные рассматриваются как результат порождающего процесса, включающего скрытые переменные [33]. В данной работе также отмечается, что порождающий процесс определяет совместное распределение вероятностей по наблюдаемым и скрытым случайным величинам. В результате это совместное распределение используется для вычисления условной вероятности скрытых переменных при наблюдаемых или апостериорной вероятностях. Выбор вероятностного подхода к моделированию обусловлен сложностью медицинских процессов и высокой вариабельностью поведения пациентов в ходе лечения.

Применение метода LDA позволяет каждому пациенту выбрать набор паттернов клинического пути с различным акцентом на значимость этих паттернов. Однако LDA выбирает одно из возможных решений, не предоставляя исследователю возможности сравнить и выбрать лучшее решение для конкретной задачи. В связи с этим ограничением был разработан альтернативный подход аддитивной регуляризации тематических моделей (ARTM), приводящей к модульности технологии тематического моделирования [34]. В данной работе для определения групп клинических путей применялась библиотека BigARTM, в основе которой лежит аддитивная регуляризация. В качестве регуляризаторов были выбраны:

- ✦ декоррелирование распределений терминов в темах для того, чтобы повысить различность этих тем;
- ✦ сглаживание распределений тем в документах;
- ✦ сглаживание распределений терминов в темах;
- ✦ разреживание распределений терминов в темах;
- ✦ разреживание распределений тем в документах.

Для оценки качества моделирования и определения оптимального количества тем используется перплексия – одна из реализованных в библиотеке BigARTM метрик. В контексте данного исследования перплексия определяет количество основных паттернов пациента в логе медицинского учреждения [13]:

$$P = \left[\exp -\frac{\sum_{\sigma\in L}\log P(e_\sigma|M)}{\sum_{\sigma\in L}|\sigma|}\right]. \qquad (12)$$

где M – модель;

e_σ – множество скрытых событий в маршруте пациента σ.

3. Вычислительный эксперимент

В качестве примера рассмотрен журнал событий голландской больницы, размещенный в открытом доступе. Журнал событий содержит 1143 маршрутов пациентов и 150291 событие. Выбор источника обусловлен тем, что базы содержат полную и открытую информацию, необходимую для исследовательских задач в области здравоохранения.

После отбрасывания аномально длинных путей верхний предел длины рассматриваемого пути полагался равным 26 событиям: $Q50 + 3 \cdot (Q75 - Q50)$, где $Q50$ – медиана, а $Q75$ соответствует 75% квантили. Для анализа данных был выбран язык программирования R. Этот язык хорошо подходит для исследовательских задач, поскольку содержит богатую библиотеку пакетов для различных сценариев [28]. Применение языка R также помогает визуализировать данные для понимания общей картины изучаемой предметной области [28]. С помощью пакета Stringdist матрица расстояний была построена методом Oca (мера расстояния Дамерау–Левенштейна).

Для поиска модели, в которой находится кластер с максимальным коэффициентом силуэта, сравнивались метод Уорда и k-medoids (*рисунок 2*), где по оси X расположены кластеры, по оси Y – значения силуэта.

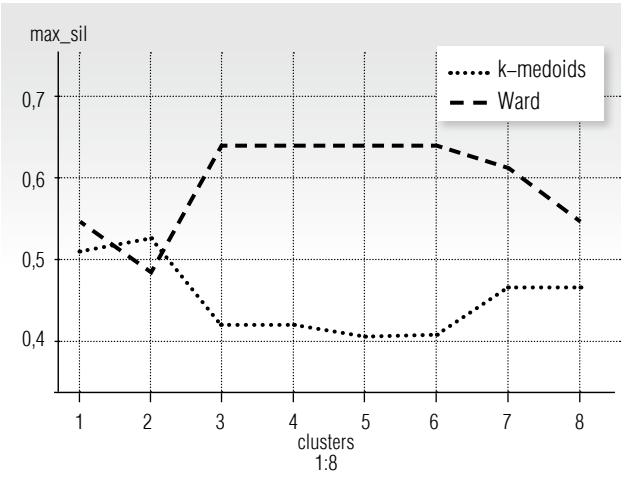

Рис. 2. Зависимость максимального коэффициента силуэта в модели от количества кластеров для методов *k*–medoids и Уорда

В соответствии с проведенным анализом значений коэффициента был выбран метод Уорда и были выявлены тенденции полученных групп (*рисунок 2*). Кластеры с низким значением были исключены (*рисунок 3*), таким образом, в результате экспериментов были выбраны кластеры 5 и 6 с наибольшим коэффициентом силуэта (*таблица 1*).

Таблица 1.

Значения коэффициента силуэта для кластеров

Номер кластера	Количество объектов	Значение коэффициента силуэта
1	118	0,02
2	239	–0,02
3	193	0,002
4	228	–0,05
5	79	0,29
6	118	0,64

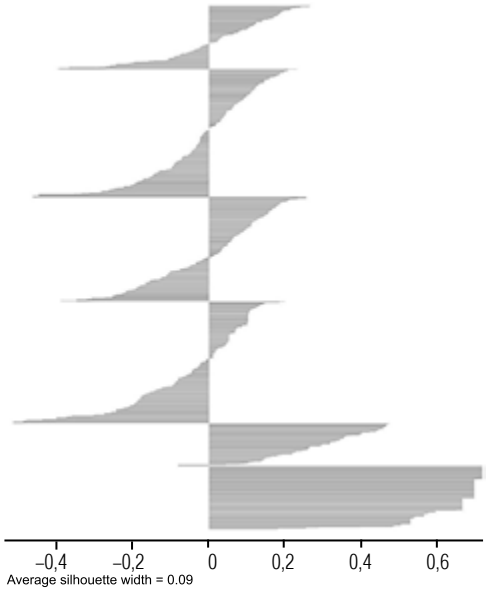

Рис. 3. Коэффициенты силуэта шести результирующих кластеров

С помощью открытого и бесплатного пакета ProcessmapR[1] была построена карта процессов исходного датасета, однако без предварительного разделения маршрутов на кластеры сложно интерпретировать полученные клинические пути.

После определения оптимального количества кластеров были отдельно сформированы процессные карты для полученных групп. Например, на процессной карте для кластера 5 (*рисунок 4*) в качестве узлов графа обозначены основные этапы клинического пути пациентов с диагнозом сепсис: старт, регистрация в соответствующем подразделении, прием антибиотиков и т.д. Дуги графа отображают переходы пациентов по этим этапам лечения, числа на дугах соответствуют количеству человек, совершающих данный переход между узлами. Более значимые траектории пациентов отмечены более широкими дугами. Таким образом, процессная карта позволяет быстро оценить наиболее загруженные маршруты медицинского учреждения. Кроме того, подобные карты могут быть интерпретированы медицинскими специалистами в разрезе сравнения их с принятыми медицинскими стандартами для выявления перегруженных ресурсных единиц и дальнейшей реорганизации процесса обслуживания.

Следующим этапом была нечеткая, или «мягкая», кластеризация исходных данных методами тематического моделирования, при которой путь пациента может относиться к нескольким шаблонам (кластерам-темам) с различными вероятностями. Латентное размещение Дирихле (latent Dirichlet allocation, LDA) [17] используется в работах по определению кластеров клинических путей [13, 32] и считается одним из стандартных методов тематического моделирования. При построении такой тематической модели возникает бесконечно много решений, ведущих к неустойчивости и плохой интерпретируемости тем [34]. Для решения подобных проблем с выбором наилучшего решения задаются дополнительные регуляризаторы, или критерии оптимальности [34]. Таким образом, возникла необходимость в разработке нового многокритериального подхода – аддитивной регуляризации тематических моделей (ARTM), предложенного в работе [34].

Применение этого более гибкого подхода для кластеризации клинических путей пациентов в исследованиях ранее не рассматривалось. В данной статье использовалась библиотека с открытым кодом BigARTM в Python, в основе которой лежит аддитивная регуляризация. Данные были преобразованы в Vowpal Wabbit формат, который принимает входные данные в определенной структуре: метка | A feature1: значение1 | B feature2: значение2. Этот формат адап-

[1] https://cran.r-project.org/web/packages/processmapR/processmapR.pdf

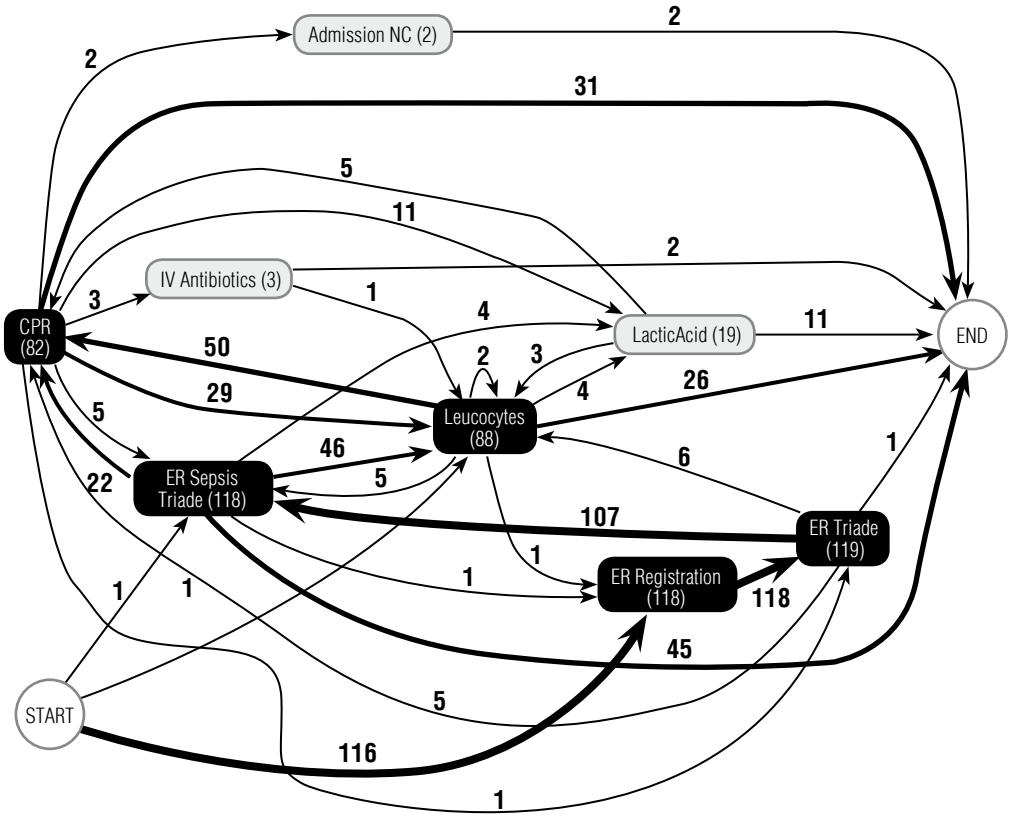

Рис 4. Процессная карта для кластера 5

тирован для разделения на категории или модальности при обучении модели. Модель была создана и обучена по начальному количеству тем $T = 300$. На основе рассчитанных параметров перплексии 63,97 и коэффициентов разреженности $\Theta = 0{,}44$ и $\Theta = 0{,}42$ было выбрано оптимальное количество кластеров, равное девяти.

Каждому уникальному пациенту была присвоена вероятностная оценка принадлежности к определенному кластеру. Например, для одного из пациентов выборки распределение по шаблонам клинических путей выглядит следующим образом: $P_1 = 0{,}017998157$, $P_2 = 0{,}059349068$, $P_4 = 0{,}5676379$, $P_6 = 0{,}35303143$. Соответственно, следующий шаг пациента с вероятностью около 57% будет соответствовать поведенческому паттерну 4 кластера.

Метод нечеткой кластеризации позволяет добавить иерархическое представление маршрутов пациентов, отображая ресурсы медицинских учреждений. Выделенные кластеры будут отправной точкой для улучшения прогноза потока прикрепленного контингента, а также для формирования рекомендаций по ресурсному оснащению больниц при развитии сервисов.

Заключение

В настоящее время медицинские учреждения располагают большими массивами данных, однако традиционный подход документирования процессов не позволяет сформировать полную картину всех траекторий пациентов и осуществлять их автоматический анализ в режиме реального времени с учетом прогнозных оценок потоков. Кроме того, разнообразная природа заболеваний отражается в высокой вариативности маршрутов.

По результатам анализа ряда исследований определены основные методы моделирования клинических путей и выявлены ограничения их применения. Представлена методология формирования групп маршрутов пациентов иерхическим агломеративным алгоритмом с методом связи Уорда, впервые для определения шаблонов клинических путей рассмотрена аддитивная регуляризацией тематических моделей (ARTM). Проведен вычислительный эксперимент на основе данных о маршрутах пациентов с диагнозом сепсис, размещенных в открытом доступе.

Полученные результаты позволяют провести предварительную оценку клинических путей пациентов

любого журнала событий, определить узкие места системы и визуализировать процессные карты деятельности медицинского учреждения.

Представленные подходы сегментации входного гетерогенного потока служат фундаментом для дальнейшей разработки имитационной модели медицинского учреждения и предоставления рекомендательных сервисов пациентам, например, чат-ботов на сайтах поликлиники по оказанию консалтинговых услуг.

Медицинские учреждения, которые первыми внедрят эти технологии, безусловно, будут иметь конкурентное преимущество. В свою очередь, руководители и другие заинтересованные стороны смогут получить доступ к полным сведениям, что позволит им принимать более обоснованные решения. ■

Литература

1. Илюшин Г.Я., Лиманский В.И. Построение системы управления потоками пациентов // Системы и средства информатики. 2015. Т. 25. № 1. С. 186–197.
2. Азанов В.Г. Структурно-функциональная модель управления потоками пациентов // Системы и средства информатики. 2016. Т. 26. № 1. С. 13–29.
3. What is a clinical pathway? Development of a definition to inform the debate / L. Kinsman [et al.] // BMC Medicine. 2010. Vol. 8. No 31. DOI: 10.1186/1741-7015-8-31.
4. Pearson S.D., Goulart-Fisher D., Lee T.H. Critical pathways as a strategy for improving care: Problems and potential // Annals of Internal Medicine. 1995. Vol. 123. No 12. P. 941–948.
5. Wakamiya S., Yamauchi K. What are the standard functions of electronic clinical pathways? // International Journal of Medical Informatics. 2009. Vol. 78. No 8. P. 543–550. DOI: 10.1016/j.ijmedinf.2009.03.003.
6. Medical guidelines presentation and comparing with electronic health record / A. Veselý [et al.] // International Journal of Medical Informatics. 2006. Vol. 75. No 3–4. P. 240–245. DOI: 10.1016/j.ijmedinf.2005.07.016.
7. Rojas E., Munoz-Gama J., Sepúlveda M., Capurro D. Process mining in healthcare: A literature review // Journal of Biomedical Informatics. 2016. No 61. P. 224–236. DOI:10.1016/j.jbi.2016.04.007.
8. Huang Z., Lu X., Duan H. On mining clinical pathway patterns from medical behaviors // Artificial Intelligence in Medicine. 2012. Vol. 56. No 1. P. 35–50. DOI: 10.1016/j.artmed.2012.06.002.
9. Rakocevic G., Djukic T., Filipovic N., Milutinović V. Computational medicine in data mining and modeling. N.Y.: Springer, 2013. DOI: 10.1007/978-1-4614-8785-2.
10. Ahmad M. A., Teredesai A., Eckert C. Interpretable machine learning in healthcare // Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI). New York, NY, USA, 4–7 June 2018. P. 447–447. DOI: 10.1109/ICHI.2018.00095.
11. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs: Cochrane database of systematic reviews and meta-analysis / T. Rotter [et al.] // Evaluation & the Health Professions. 2010. Vol. 35. No 1. P. 3–27. DOI: 10.1177/0163278711407313.
12. Prodel M. Process discovery, analysis and simulation of clinical pathways using health-care data / Université de Lyon, 2017. [Электронный ресурс]: https://tel.archives-ouvertes.fr/tel-01665163/document (дата обращения 25.11.2019).
13. Discovery of clinical pathway patterns from event logs using probabilistic topic models / Z. Huang [et al.] // Journal of Biomedical Informatics. 2014. No 47. P. 39–57. DOI: 10.1016/j.jbi.2013.09.003.
14. Lin F., Chou S., Pan S., Chen Y. Mining time dependency patterns in clinical pathways // International Journal of Medical Informatics. 2001. Vol. 62. No 1. P. 11–25. DOI: 10.1016/S1386-5056(01)00126-5.
15. Cote M.J., Stein W.E. A stochastic model for a visit to the doctor's office // Mathematical and Computer Modelling. 2007. Vol. 45. No 3–4. P. 309–323. DOI: 10.1016/j.mcm.2006.03.022.
16. Zhang Y., Padman R., Patel N. Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data // Journal of Biomedical Informatics. 2015. No 58. P. 186–197.
17. Blei D.M., Ng A.Y. Jordan M.I. Latent Dirichlet allocation // The Journal of Machine Learning Research. 2003. No 3. P. 993–1022.
18. Fernández-Llatas C., Benedi J.-M., García-Gómez J.M., Traver V. Process mining for individualized behavior modeling using wireless tracking in nursing homes // Sensors (Basel). 2013. Vol. 13. No 11. P. 15434–15451. DOI: 10.3390/s131115434.
19. van der Aalst W.M.P. Process mining: Discovery, conformance and enhancement of business processes. Springer, 2011. DOI: 10.1007/978-3-642-19345-3.
20. van der Aalst W.M.P. Process mining and simulation: A match made in heaven! // Proceedings of the 50th Computer Simulation Conference (SummerSim 2018). Bordeaux, France, 9–12 July 2018. DOI: 10.22360/summersim.2018.scsc.005.
21. van der Aalst W.M.P. Process mining: Data science in action. Berlin: Springer-Verlag, 2016.
22. van der Aalst W.M.P. Process mining manifesto // Business Process Management Workshops. Springer, 2011. P. 169–194. DOI: 10.1007/978-3-642-28108-2_19.
23. Fern ndez-Llatas C., Meneu T., Bened J.M., Traver V. Activity-based process mining for clinical pathways computer aided design // Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. Buenos Aires, Argentina, 31 August – 4 September 2010. P. 6178–6181. DOI: 10.1109/IEMBS.2010.5627760.
24. Kovalchuk S.V., Funkner A.A., Metsker O.G., Yakovlev A.N. Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification // Journal of Biomedical Informatics. 2018. No 82. P. 128–142.

25. Williams R., Buchan I., Prosperi M., Ainsworth J. Using string metrics to identify patient journeys through care pathways // Proceedings of the AMIA Annual Symposium, Washington, DC, USA, 15–19 November 2014. P. 1208–1217.

26. Kaufmann L., Rousseeuw P. Clustering by means of medoids // Data analysis based on the L1-norm and related methods. 1987. P. 405–416.

27. Ward J.H. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. 1963. Vol. 58. No 301. P. 236–244.

28. Зайцев Р.Д., Бритков В.Б. Применение языка R для многомерной кластеризации временных рядов с целью анализа динамики научно-технического развития // Труды Второй молодежной научной конференции «Задачи современной информатики». Москва, 29–30 октября 2015 г. М.: ФИЦ ИУ РАН. С. 92–98.

29. Ferreira L., Hitchcock D. A comparison of hierarchical methods for clustering functional data // Communications in Statistics – Simulation and Computation. 2009. No 38. P. 1925–1949. DOI: 10.1080/03610910903168603.

30. Коннов И.В., Кашина О.А., Гильманова Э.И. Решение задачи кластеризации методами оптимизации на графах // Ученые записки Казанского университета. Сер. Физико-математические науки. 2019. Т. 161. Кн. 3. С. 423–437. DOI: 10.26907/2541-7746.2019.3.423-437.

31. Boytsov L. Indexing methods for approximate dictionary searching // Journal of Experimental Algorithmics. 2011. Vol. 16. No 1. Article No 1.1. DOI: 10.1145/1963190.1963191.

32. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. No 20. P. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

33. Huang Z., Lu X., Duan H., Fan W. Summarizing clinical pathways from event logs // Journal of Biomedical Informatics. 2013. Vol. 46. No 1. P. 111–127. DOI: 10.1016/j.jbi.2012.10.001.

34. Vorontsov K.V., Potapenko A.A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. Springer, 2014. P. 265–267.

Об авторах

Прокофьева Елизавета Сергеевна

аспирант, кафедра инноваций и бизнеса в сфере информационных технологий,
Национальный исследовательский университет «Высшая школа экономики»,
101000, г. Москва, ул. Мясницкая, д. 20;

E-mail: prokofyeva.liza@gmail.com

ORCID: 0000-0003-1322-2932

Зайцев Роман Дмитриевич

старший эксперт по анализу данных, ГК «ФОРС»,
129272, г. Москва, ул. Трифоновский тупик, д. 3;

E-mail: roman.zaitsev@fors.ru

ORCID: 0000-0002-8313-3727

Clinical pathways analysis of patients in medical institutions based on hard and fuzzy clustering methods

Elizaveta S. Prokofyeva[a]
E-mail: prokofyeva.liza@gmail.com

Roman D. Zaytsev[b]
E-mail: Roman.Zaitsev@fors.ru

[a] National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

[b] FORS Group
Address: 3, Trifonovskiy Tupik Street, Moscow 129272, Russia

Abstract

Modeling the processes in a healthcare system plays a large role in understanding its activities and serves as the basis for increasing the efficiency of medical institutions. The tasks of analyzing and modeling large amounts of urban healthcare data using machine learning methods are of particular importance and relevance for the development of industry solutions in the framework of digitalization of the economy, where data is the key factor in production. The problem of automatic analysis and determination of clinical pathways groups of patients based on clustering methods is considered in this research. Existing projects in this area reflect a great interest on the part of the scientific community in such studies; however, there is a need to develop a number of methodological approaches for their further practical application in urban outpatient institutions, taking into account the specifics of the organization being analyzed. The aim of the study is to improve the quality of management and segmentation of patient input flow in urban medical institutions based on cluster analysis methods for the further development of recommendation services. One approach to achieving this goal is the development and implementation of clinical pathways, or patient trajectories. In general, the clinical pathway of a patient might be interpreted as the trajectory when receiving medical services in respective institutions. The approach of developing groups of patient routes by the hierarchical agglomerative algorithm with the Ward method and Additive Regularization of Topic Models (ARTM) is presented in this article. A computational experiment based on public data on the routes of patients with a diagnosis of sepsis is described. One feature of the proposed approach is not just the automation of the determination of similar groups of patient trajectories, but also the consideration of clinical pathways patterns to form recommendations for organizing the resource allocation of a medical institution. The proposed approach to segmenting the input heterogeneous flow of patients in urban medical institutions on the basis of clustering consists of the following steps: 1) preparing the data of the medical institution in the format of an event log; 2) encoding patient routes; 3) determination of the upper limit of the clinical pathway length; 4) hierarchical agglomerative clustering; 5) additive regularization of topic models (ARTM); 6) identifying popular patient route patterns. The resulting clusters of routes serve as the foundation for the further development of a simulation model of a medical institution and provide recommendations to patients. In addition, these groups may underlie the development of the robotic process automation system (RPA), which simulates human actions and allows you to automate the interpretation of data to manage the resources of the institution.

Key words: cluster analysis; data; hierarchical clustering; topic modeling; silhouette coefficient; healthcare; clinical pathways; process mining.

Citation: Prokofyeva E.S., Zaytsev R.D. (2020) Clinical pathways analysis of patients in medical institutions based on hard and fuzzy clustering methods. *Business Informatics*, vol. 14, no 1, pp. 19–31.
DOI: 10.17323/2587-814X.2020.1.19.31

References

1. Ilyushin G.Ya., Limanskij V.I. (2015) Development of the patient flow management system. *Systems and Approaches of Informatics*, vol. 25, no 1, pp. 186–197 (in Russian).
2. Azanov V.G. (2016) Structural-functional model of patient flow management. *Systems and Approaches of Informatics*, vol. 26, no 1, pp. 13–29 (in Russian).
3. Kinsman L., Rotter T., James E., Snow P., Willis J. (2010) What is a clinical pathway? Development of a definition to inform the debate. *BMC Medicine*, vol. 8, no 31. DOI: 10.1186/1741-7015-8-31.
4. Pearson S.D., Goulart-Fisher D., Lee T.H. (1995) Critical pathways as a strategy for improving care: Problems and potential. *Annals of Internal Medicine*, vol. 123, no 12, pp. 941–948.
5. Wakamiya S., Yamauchi K. (2009) What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*, vol. 78, no 8, pp. 543–550. DOI: 10.1016/j.ijmedinf.2009.03.003.
6. Veselý A., Zvárová J., Peleska J., Buchtela D., Anger Z. (2006) Medical guidelines presentation and comparing with electronic health record. *International Journal of Medical Informatics*, vol. 75, no 3–4, pp. 240–245. DOI: 10.1016/j.ijmedinf.2005.07.016.
7. Rojas E., Munoz-Gama J., Sep lveda M., Capurro D. (2016) Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, no 61, pp. 224–236. DOI:10.1016/j.jbi.2016.04.007.
8. Huang Z., Lu X., Duan H. (2012) On mining clinical pathway patterns from medical behaviors. *Artificial Intelligence in Medicine*, vol. 56, no 1, pp. 35–50. DOI: 10.1016/j.artmed.2012.06.002.
9. Rakocevic G., Djukic T., Filipovic N., Milutinović V. (2013) *Computational medicine in data mining and modeling.* N.Y.: Springer. DOI: 10.1007/978-1-4614-8785-2.
10. Ahmad M. A., Teredesai A., Eckert C. (2018) Interpretable machine learning in healthcare. Proceedings of the *2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY, USA, 4–7 June 2018*, pp. 447–447. DOI: 10.1109/ICHI.2018.00095.
11. Rotter T., Kinsman L., James E.L., Machotta A., Gothe H., Willis J., Snow P., Kugler J. (2010) Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs: Cochrane database of systematic reviews and meta-analysis. *Evaluation & the Health Professions*, vol. 35, no 1, pp. 3–27. DOI: 10.1177/0163278711407313.
12. Prodel M. (2017) *Process discovery, analysis and simulation of clinical pathways using health-care data.* Université de Lyon. Available at: https://tel.archives-ouvertes.fr/tel-01665163/document (accessed 25 November 2019).

13. Huang Z., Dong W., Ji L., Gan C., Lu X., Duan H. (2014) Discovery of clinical pathway patterns from event logs using probabilistic topic models. *Journal of Biomedical Informatics*, no 47, pp. 39–57. DOI: 10.1016/j.jbi.2013.09.003.

14. Lin F., Chou S., Pan S., Chen Y. (2001) Mining time dependency patterns in clinical pathways. *International Journal of Medical Informatics*, vol. 62, no 1, pp. 11–25. DOI: 10.1016/S1386-5056(01)00126-5.

15. Cote M.J., Stein W.E. (2007) A stochastic model for a visit to the doctor's office. *Mathematical and Computer Modelling*, vol. 45, no 3–4, pp. 309–323. DOI: 10.1016/j.mcm.2006.03.022.

16. Zhang Y., Padman R., Patel N. (2015) Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of Biomedical Informatics*, no 58, pp. 186–197.

17. Blei D.M., Ng A.Y. Jordan M.I. (2003) Latent Dirichlet allocation. *The Journal of Machine Learning Research*, no 3, pp. 993–1022.

18. Fernández-Llatas C., Benedi J.-M., García-Gómez J.M., Traver V. (2013) Process mining for individualized behavior modeling using wireless tracking in nursing homes. *Sensors (Basel)*, vol. 13, no 11, pp. 15434–15451. DOI: 10.3390/s131115434.

19. van der Aalst W.M.P. (2011) *Process mining: Discovery, conformance and enhancement of business processes*. Springer. DOI: 10.1007/978-3-642-19345-3.

20. van der Aalst W.M.P. (2018) Process mining and simulation: A match made in heaven! Proceedings of the *50th Computer Simulation Conference (SummerSim 2018). Bordeaux, France, 9–12 July 2018*. DOI: 10.22360/summersim.2018.scsc.005.

21. van der Aalst W.M.P. (2016) *Process mining: Data science in action*. Berlin: Springer-Verlag.

22. van der Aalst W.M.P. (2011) Process mining manifesto. *Business Process Management Workshops*. Springer, pp. 169–194. DOI: 10.1007/978-3-642-28108-2_19.

23. Fern ndez-Llatas C., Meneu T., Bened J.M., Traver V. (2010) Activity-based process mining for clinical pathways computer aided design. Proceedings of the *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. Buenos Aires, Argentina, 31 August – 4 September 2010*, pp. 6178–6181. DOI: 10.1109/IEMBS.2010.5627760.

24. Kovalchuk S.V., Funkner A.A., Metsker O.G., Yakovlev A.N. (2018) Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification. *Journal of Biomedical Informatics*, no 82, pp. 128–142.

25. Williams R., Buchan I., Prosperi M., Ainsworth J. (2014) Using string metrics to identify patient journeys through care pathways. Proceedings of the *AMIA Annual Symposium, Washington, DC, USA, 15–19 November 2014*, pp. 1208–1217.

26. Kaufmann L., Rousseeuw P. (1987) Clustering by means of medoids. *Data analysis based on the L1-norm and related methods*, pp. 405–416.

27. Ward J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, vol. 58, no 301, pp. 236–244.

28. Zaitsev R.D., Britkov V.B. (2015) The use of the R language for multidimensional clustering of time series in order to analyze the dynamics of scientific and technological development. *Transactions of the Second Youth Scientific Conference "Problems of Modern Computer Science", Moscow, 29–30 October 2015*, pp. 92–98.

29. Ferreira L., Hitchcock D. (2009) A comparison of hierarchical methods for clustering functional data. *Communications in Statistics – Simulation and Computation*, no 38, pp. 1925–1949. DOI: 10.1080/03610910903168603.

30. Konnov I.V., Kashina O.A., Gilmanova E.I. (2019) Solving the clustering problem by optimization methods on graphs. *Scientific Letters of the Kazan University, Series Physical and Mathematical Sciences*, vol. 161, pp. 423–437 (in Russian). DOI: 10.26907/2541-7746.2019.3.423-437.

31. Boytsov L. (2011) Indexing methods for approximate dictionary searching. *Journal of Experimental Algorithmics*, vol. 16, no 1, article no 1.1. DOI: 10.1145/1963190.1963191.

32. Rousseeuw P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

33. Huang Z., Lu X., Duan H., Fan W. (2013) Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, vol. 46, no 1, pp. 111–127. DOI: 10.1016/j.jbi.2012.10.001.

34. Vorontsov K.V., Potapenko A.A. (2014) Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *AIST'2014, Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science*. Springer, pp. 265–267.

About the authors

Elizaveta S. Prokofyeva
Doctoral Student, Department of Innovation and Business in Information Technologies,
National Research University Higher School of Economics, 20, Myasnitskaya Street, Moscow 101000, Russia;
E-mail: prokofyeva.liza@gmail.com
ORCID: 0000-0003-1322-2932

Roman D. Zaytsev
Senior Expert for Data Analysis, FORS Group, 3, Trifonovskiy Tupik Street, 129272 Moscow, Russia;
E-mail: roman.zaitsev@fors.ru
ORCID: 0000-0002-8313-3727