Эволюционные подходы в задачах моделирования клинических путей с использованием статической и динамической идентификации моделей

А. А. Функнер[1], И. О. Кисляковский[2],
О. Г. Мецкер[3]
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
[1]funkner.anastasia@gmail.com, [2]kisliakovskiii@niuitmo.ru,
[3]olegmetsker@gmail.com

А. Н. Яковлев
ФГБУ «НМИЦ им. В. А. Алмазова» Минздрава России
yakovlev_an@almazovcentre.ru

***Аннотация*. В работе рассматривается применение эволюционных подходов к задачам идентификации моделей клинических путей по данным электронных медицинских карт пациентов, проходящих лечение в специализированном медицинском центре. Предлагается двухуровневый гибридный подход, позволяющий, с одной стороны идентифицировать типовые клинические пути с использованием кластеризации, а с другой стороны, реализовать процедуры эволюционного усвоения данных в предсказательные модели клинических путей для предсказания развития состояния пациентов по мере накопления данных электронных медицинских карт. Рассматриваются примеры применения данного подхода в рамках задачи моделирования клинических путей пациентов с острым коронарным синдромом (ОКС).**

Ключевые слова: эволюционные вычисления; гибридное моделирование; усвоение данных; клинические пути; острый коронарный синдром

I. Введение

Лечение пациента – это комплексный и слабоструктурированный процесс. Множество факторов, влияющих на ход лечения, остаются неизвестными. В условиях такой высокой неопределенности сложно разработать модель, достаточно точно описывающую необходимые медицинские процессы. Эволюционные, в том числе генетические, алгоритмы позволяют исследовать пространство структур процессов на основе имеющихся данных.

II. Моделирование слабоструктурированных процессов

При моделировании слабоструктурированных процессов генетические алгоритмы могут быть использованы дважды. На рис. 1 описаны этапы эволюции пространства возможных структур модели и эволюции популяции моделей.

Работа выполнена при финансовой поддержке РНФ, проект №14-11-00823

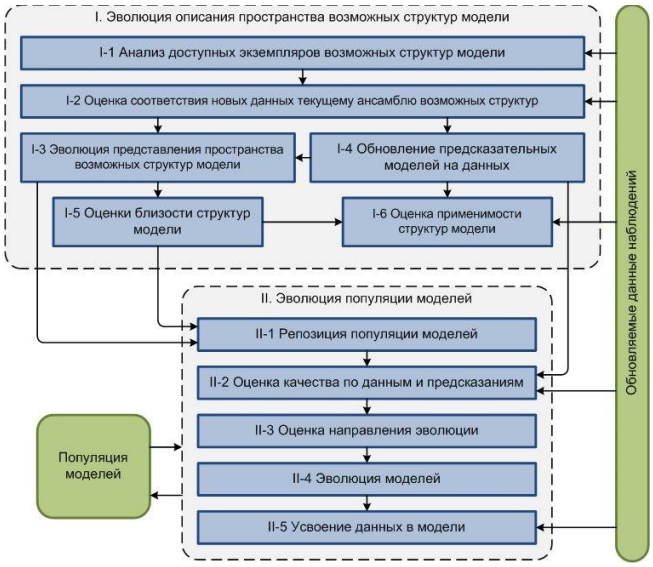

Рис. 1. Общая схема эволюционного моделирования

Моделирование процессов происходит после определения пространства возможных структур процессов и моделей. В ходе исследования пространства структур и моделирования могут поступать новые данные о процессе. В любом случае необходимо проверить насколько новые данные соответствуют уже имеющимся и при необходимости изменять структуру пространства и корректировать предсказания, сделанные на основе прежних данных и с помощью построенных моделей. Такую идентификацию моделей называют статической.

После определения всевозможных структур процессов и моделей можно приступать к предсказаниям в реальном времени. Если данные поступают в режиме реального времени и заведомо невозможно знать о всевозможных структурах процессов, то необходимо корректировать не только само пространство структур процессов и моделей, но и подбирать наилучшую модель. Благодаря усвоению данных можно изменять параметры подобранной модели в

режиме реального времени. Такой тип идентификации модели принято называть динамическим.

Далее рассматривается пример применения концепции эволюционного моделирования для определения и предсказания структуры клинических путей пациентов с острым коронарным синдромом.

III. Идентификация клинических путей

В данном разделе представлен метод идентификации клинических путей пациентов. ФГБУ «НМИЦ им. В.А. Алмазова» Минздрава России предоставил 3434 электронные медицинские карты пациентов с диагнозом поступления острый коронарный синдром (ОКС) с 2010 по 2015 года. Электронная медицинская карта содержит описание событий, которые происходили с пациентов во время его госпитализации в ЛПУ (поступление в стационар, взятие анализов, осмотры врачом, операции, переводы из отделения в отделение). Для пациентов с сердечно-сосудистыми заболеваниями можно выделить группу ключевых отделений в медицинском центре: отделение приемного покоя (ПО), кардиологическое отделение №1 и №2 (КО №1 и КО №2), отделение рентгенохирургических методов диагностики и лечения №1 И №2 (РХМДиЛ №1 и РХМДиЛ №2), отделение анестезиологии и реанимации №1 и №2 (ОАиР №1 и ОАиР №2). На основе медицинских электронных карт для каждого пациента можно составить его последовательность переводов из отделения в отделения. Можно предположить, что пациент перемещается в определенном порядке из отделения в отделение вследствие тяжести его состояния, особенностей лечения и скорости восстановления. Все рассматриваемые пациенты имеют одинаковые диагнозы, однако значительно отличаются по последовательности и времени пребывания в разных отделениях.

Последовательность перемещений для каждого пациента кодируется последовательностью символов. Таким образом, каждому пациенту в соответствие ставится символьная строка. Несмотря на то, что все пациенты имеют схожие заболевания и имеют один и тот же диагноз поступления, острый коронарный синдром, каждый из них имеет свою историю болезни, разные сопутствующие и конкурирующие заболевания, а также определённую тяжесть основного заболевания на момент поступления в стационар. Можно предположить, что пациенты значительно отличаются друг от друга, что сказывается на их перемещениях между отделениями. Поэтому следующим этапом идентификации клинических путей стало разделение пациентов на группы с помощью алгоритмов кластеризации. Для этого был использован метод k-средних, который подходит для работы с большими объемами данных. Для определения различия между разными последовательностями отделений было выбрано расстояние Левенштайна, которое позволяет определять расстояние между строками разной длины. Число кластеров определялось с помощью внутри- и межкластерного расстояния [1]. Итак, было выделено 10 кластеров.

Далее необходимо определить основные паттерны поведения пациентов внутри каждого кластера: нужно обобщить все пути пациентов. Для этого с помощью генетического алгоритма выращивается шаблон для каждого кластера, затем последовательности выравниваются согласно этому шаблону, что даёт возможность визуализировать клинические пути пациентов.

A. *Применение генетического алгоритма для поиска наилучшего шаблона*

Шаблон является обобщением множества последовательностей состояний и ищется, для того чтобы в общем описать все последовательности состояний. Таким образом, найденный шаблон будет включать типичные клинические пути для данного множества последовательностей состояний.

Очевидно, что для любого множества можно подобрать такой шаблон, под который выровняться все последовательности. Однако, такой шаблон будет состоять из большого количества состояний и не позволит понять особенности последовательностей перемещений. Поэтому необходимо подобрать такой шаблон, под который смогут выровняться наибольшее число последовательной и в то же время шаблон будет иметь наименьшую длину. Сформулированная задача называется проблемой поиска кратчайшей общей надстройки и не решается за полиномиальное время [2]. В данном случае, для нахождение наилучшего шаблона для определенного множества последовательностей состояний можно решить задачу многокритериальной оптимизации. Для решения этой задачи используем генетический алгоритм, так как он подходит для поиска глобального экстремума и его можно модифицировать для решения задач многокритериальной оптимизации.

Определим функция выравнивания V = V(T, B), которая по выбранному шаблону $T \in \Omega$ и множеству последовательностей B определяет количество последовательностей, которое нельзя выровнять по данному шаблону T. Определим функцию длины шаблона L = L(T), которая выбранному шаблону $T \in \Omega$ ставит в соответствие его длину. Тогда функции V и L являются целевыми, и задача многокритериальной оптимизации формулируется следующем образом:

$$\min_{T \in \Omega} \{V(T, S), L(T)\}$$

Далее с помощью генетического алгоритма найдём решения (1) оптимальные по Парето. Решение $\dot{T} \in \Omega$ называется оптимальным по Парето, если не существует $T \in \Omega$ такого, что $f_i(T) \leq f_i(\dot{T})$ для всех $i = 1, \dots, k$ и $f_i(T) < f_i(\dot{T})$ для хотя бы одного i. Функции f_i – это целевые функции. Также данное определение сформулировано для задачи минимизации всех целевых функций. Множество оптимальных по Парето решений называется фронтом Парето и обозначается как P(Ω). Целевой вектор является оптимальным по Парето, если соответствующий ему вектор из области определения

также оптимален по Парето. Множество оптимальных по Парето целевых векторов можно обозначить как P(Z) [3].

Для генетического алгоритма необходимо определить вид генотипов, функции приспособленности и способы скрещивание, мутации и селекции генотипов. В нашем случае генотипами являются шаблонами. Начальная популяция генотипов формируется случайным образом на основе всех состояний, которые встречаются в множестве B. Однако необходимо ограничить размеры генотипов в начальной популяции. Функция приспособленности в данном случае является вектор-функцией и состоит из двух компонентов: функции выравнивания и функции длины.

Селекция происходит на основе определения фронта Парето: решения лежащие на фронте Парето определяются как самые приспособленные, становятся родителями и остаются в популяции вместе с новым поколением (рис. 2). Если фронт Парето состоит из недостаточного количества решений, а генотипов-родителей нужно больше, то их можно добрать из остальной популяции случайным образом. Скрещивание и получение генотипов нового поколения происходит на основе кроссинговера [4]. Мутации генотипов происходят по тому же принципу, что и в геномах живых организмов. Используется три вида мутации: инсерция, делеция и замена [5]. В данный момент условие остановки генетического алгоритма не выбрано, поэтому изначально задаётся количество поколений. Результат работы генетического алгоритма является множество шаблонов, являющихся фронтом Парето P(Ω) в последнем поколении.

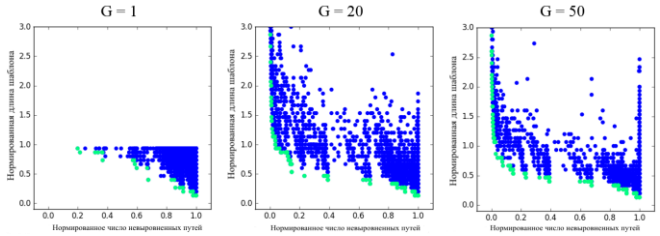

Рис. 2. Фронт Парето для поколений 1, 20 и 50. Фронт Парето содержит оптимальные шаблоны для выравнивания клинических путей всех пациентов

Получив решения из фронта Парето, необходимо выбрать наиболее значимые из них. В [6] Chaudari и др. предлагают выбирать значимые решения на основе кластеризации.

После определения шаблона клинические пути кластера могут быть визуализированы с помощью графа, где вершины – это отделения медицинского центра, ребра – это возможные переводы между отделениями, а веса ребер – это количество пациентов, переведенных таким образом между отделениями. На рис. 3 показаны три наиболее многочисленных кластера. Графовая визуализация легко интерпретировать полученные кластеры.

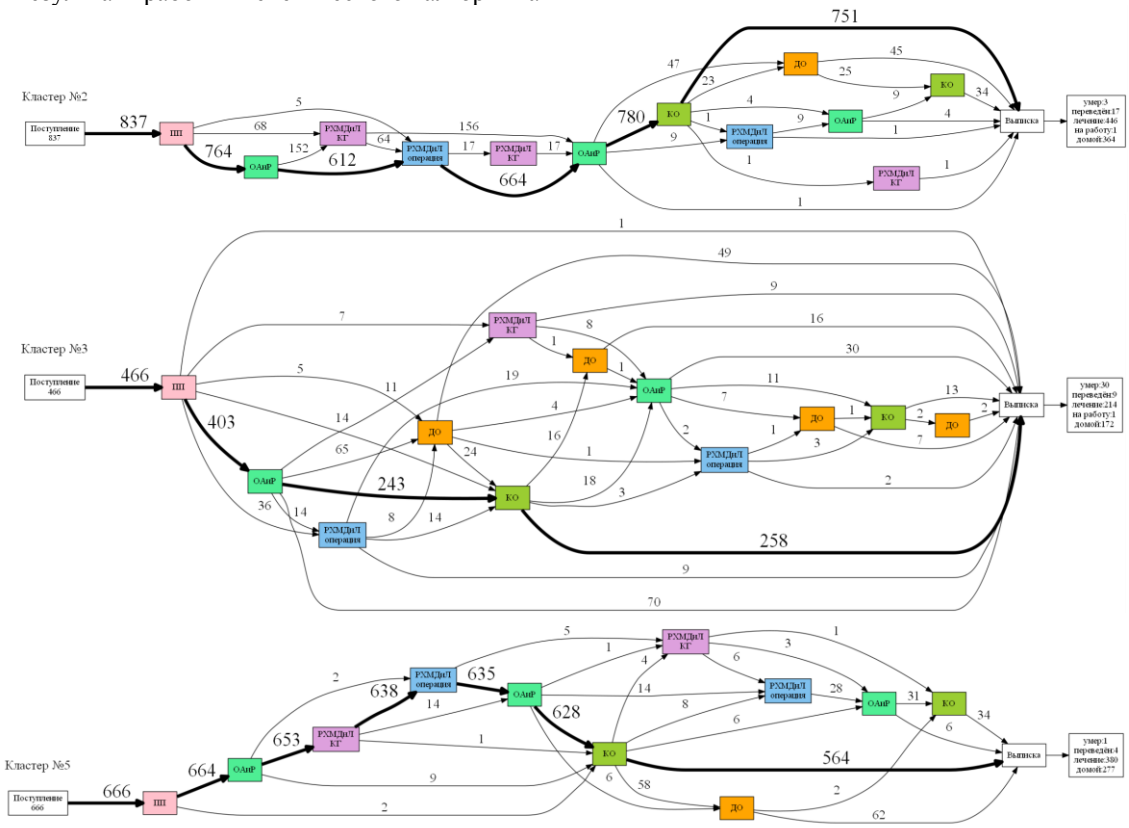

Рис. 3. Наиболее многочисленные кластера №2, №3 и №5, полученные с помощью эволюционных шаблонов клинических путей. Обозначения: ПП – приемный покой, ОАиР – отделение анестезиологии и реанимации, КО – кардиологическое отделение, РХМДиЛ – Отделение рентгенохирургических методов диагностики и лечения, КГ – коронарография, ДО – другие отделения медицинского центра

IV. Эволюционный алгоритм поиска позиции модели процесса

В рамках исследования эволюции популяции моделей и усвоения данных наблюдений был реализован эволюционный алгоритм поиска позиции модели процесса в пространстве возможных моделей процесса в графовом представлении. Сначала граф состоит из наблюдаемых клинических путей, дополненных синтетическими путями, распределение кластеров среди которых соответствует выделенным кластерам на прошлом шаге. Далее, используя клинический путь, не включенный в граф попытаемся предсказать по уже известным отделениям пациента его последующие перемещения. По мере накопления данных популяция предсказанных путей локализуется вокруг целевой позиции в графе, что соответствует эволюции популяции моделей (рис. 1). При этом на промежуточных этапах становится возможной оценка потенциальных путей развития эпизода, то есть алгоритм демонстрирует на каком шаге имеющейся информации будет достаточно, чтобы определить оставшуюся часть клинического пути. Видео-демонстрация процесса эволюции доступна по адресу https://rnf.escience.ifmo.ru/14-11-00823/.

На рис. 4 в виде графа представлены клинические пути пациентов и близость (похожесть) данных клинических путей. Сами клинические пути представлены в виде вершин графа: фиолетовым цветом отмечены реальные клинические пути, оранжевым цветом отмечены пути, сгенерированные на основе данных о распределении клинических путей по кластерам, зелёным цветом отмечена ключевая вершина, с помощью которой выполняется демонстрация работы алгоритма. Эволюционный алгоритм, получая на каждом шаге дополнительную информацию о ключевой вершине, генерирует новые клинические пути, соответствующие вершины которых концентрируются вокруг всё меньшего количества областей исходного графа, постепенно сходясь к области вокруг ключевой вершины.

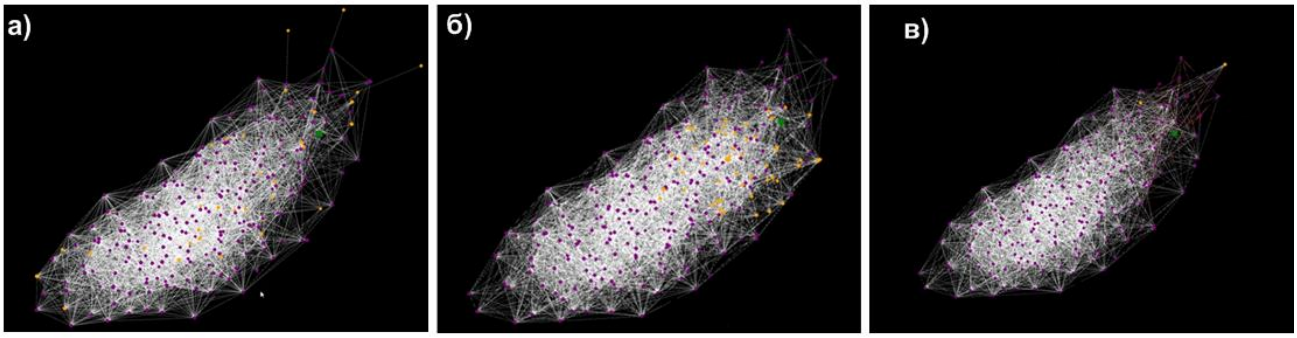

Рис. 4. Визуализация состояния графа клинических путей а) в начале процесса эволюции; б) на промежуточном шаге; в) на конечном шаге

V. Заключение

Предложенный подход к моделированию слабоструктурированных процессов может быть применен к любым задачам, когда сложно или невозможно сделать прямые выводы о структуре системы и ее закономерностях в силу ее ненаблюдаемости, изменчивости или сложности структуры.

Текущие исследование ещё не окончено. В будущем планируется разработать эволюционные модели на данных пациентов с хроническими заболеваниями, а также выявить паттерны поведения пользователей социальных сетей и частных клиентов банков.

Описанные модели идентификации клинических путей и их предсказания планируется внедрить в разработанную ранее систему поддержки принятия решений для больных с ОКС [7].

Список литературы

[1] Ray S., Turi R.H. Determination of number of clusters in k-means clustering and application in colour image segmentation // Proc. 4th Int. Conf. Adv. pattern Recognit. Digit. Tech. 1999. P. 137–143.

[2] Yu Y.W. Approximation hardness of Shortest Common Superstring variants // arXiv. 2016. № Table 1. P. 1–10.

[3] Luc D.T. Multi-Objective Linear Programming Problem. Avignon, France: Springer International Publishing, 2016. Vol. 3, № 1. 31-38 p.

[4] Brownlee J. Clever Algorithms: Nature-inspired Programming Recipes. Jason Brownlee, 2011. 436 p.

[5] DNA and Mutations [Electronic resource]. URL: http://evolution.berkeley.edu/evolibrary/article/mutations_03.

[6] Chaudhari P., Dharaskar R., Thakare V.M. Computing the Most Significant Solution from Pareto Front obtained in Multi-objective Evolutionary // Ijacsa. 2010. Vol. 1, № 4. P. 63–68.

[7] Krikunov A.V. et al. Complex data-driven predictive modeling in personalized clinical decision support for Acute Coronary Syndrome episodes // Procedia Comput. Sci. Elsevier Masson SAS, 2016. Vol. 80. P. 518–529.