Алгоритм оценки продолжительности стадий заболевания по набору неполных данных

П.Н. Подзолков
Тюменский государственный университет
Тюмень, Россия
p.n.podzolkov@utmn.ru

***Аннотация*—В статье рассматриваются вопросы математического моделирования развития заболеваний. Поднимается проблема оценки временных показателей по неполным данным для построения моделей многостадийных процессов. Предложен алгоритм выравнивания набора последовательностей, содержащих неполною информацию о течении заболевания, для оценки продолжительностей его стадий. Представлены результаты тестирования алгоритма на симулированных данных.**

Ключевые слова— моделирование заболеваний, выравнивание, неполные данные, компартментальная модель эпидемии, цепь Маркова.

1. Введение

Хронические и инфекционные заболевания до сих пор остаются значительной проблемой мирового здравоохранения. Существуют различные методы математического моделирования и прогнозирования развития заболеваний. Большинство методов основано на представлении заболевания в виде многостадийного процесса, проходя через который, пациент переходит последовательно из одной стадии в другую. В соответствии с работой Tolles J. и Luong T. B. по такому механизму работают компартментальные модели эпидемий. Андреев Д. М. показал, что модели хронических заболеваний, применяющие цепи Маркова, основаны на том же принципе.

Одним из основных этапов построения подобных моделей является вычисление средней продолжительности каждой стадии и соответственно оценка вероятности перехода в следующую стадию в каждый момент времени. Такие показатели могут быть вычислены по результатам наблюдений за пациентами, проходившими через процесс данного заболевания. Однако в большинстве случаев есть информация только о нескольких, проведённых с различными интервалами обследованиях пациента. Согласно Михальскому А. И. подобная ситуация соответствует интервально-цензурированным данным. Результатом каждого обследования является определённая стадия заболевания, однако в промежутках между обследованиями данных о состоянии индивида нет и, следовательно, момент перехода в следующее состояние, если таковой произошёл, не известен.

Не имея установленных моментов переходов между стадиями, нельзя определить длительность протекания каждой стадии заболевания и соответственно оценить вероятность перехода из неё в конкретный момент времени. Для решения этой проблемы предлагается алгоритм выравнивания последовательностей, описывающих прохождение заболевания у индивидов.

2. Критерии алгоритма

На вход алгоритму подаются строки, содержащие информацию о пройденных состояниях, разделённых интервалами различной длины, которые соответствуют промежуткам времени между обследованиями. Задачей алгоритма является определение смещения для каждой последовательности, такое, чтобы моменты наступления и завершения стадий максимально соответствовали другим последовательностям.

Для определения критериев алгоритма были введены следующие понятия. Гарантированный промежуток стадии (ГПС) – промежуток времени в последовательности наблюдений, который начинается с первым вхождением данной стадии и заканчивается последним вхождением данной стадии. Допустимый промежуток стадии (ДПС) – промежуток времени в последовательности наблюдений, который начинается после завершения ГПС предыдущей наблюдаемой стадии и заканчивается перед ГПС следующей наблюдаемой стадии. Примеры определения указанных промежутков выделены на Рис. 1.

Критерием качества выравнивания в таком случае становится максимальное соответствие ДПС и ГПС каждой стадии у последовательности с данными промежутками соответствующей стадии другой последовательности.

ГПС стадии b	a	-	b	-	-	b	-	c
ДПС стадии b	a	-	b	-	-	b	-	c

Рис. 1. Пример последовательности с выделенными серым цветом промежутками стадии b

3. Парное выравнивание

Для определения качества конкретного взаимного расположения двух последовательностей вводится штраф за расстояние между серединами ДПС и ГПС соответствующих стадий двух последовательностей.

Во многих случаях для конкретной стадии нельзя определить границы ДПС или ГПС. В случае, если нет информации о предыдущих или последующих состояниях пациента, нельзя определить ДПС выбранной стадии. А если стадия ни разу не наблюдалась у пациента, то нельзя определить её ГПС. Таким образом, необходим набор правил, по которым будет находится штраф между двумя выравниваемыми последовательностями. В таблице 1 указан вариант правил для оценки штрафа расстояний. В случае, если у одной из последовательностей невозможно определить ни один промежуток для выбранной стадии, то эта стадия не влияет на итоговый штраф выравнивания выбранной пары последовательностей. Для вычисления штрафа каждой стадии необходимо найти сумму квадратов расстояний, указанных в Таблице 1. Штраф каждой

стадии прибавляется к общему счётчику штрафа выравнивания двух последовательностей. Таким образом, для каждого варианта смещения двух последовательностей вычисляется величина характеризующая неудовлетворительность данного варианта. Варианты с наименьшим штрафом выбираются как наиболее корректные.

ТАБЛИЦА 1. ПРАВИЛА ВЫЧИСЛЕНИЯ РАССТОЯНИЙ МЕЖДУ ЦЕНТРАМИ ПРОМЕЖУТКОВ КОНКРЕТНОЙ СТАДИИ В ДВУХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

Интервалы, определённые во 2ой последовательности	**Интервалы, определённые в 1ой последовательности**		
	ГПС	***ДПС***	***ГПС и ДПС***
ГПС	$dist^a(g_1^b, g_2^c)$	$dist(d_1^d, g_2)$	$dist(g_1, g_2)$
ДПС	$dist(g_1, d_2^e)$	$dist(d_1, d_2)^c$	$dist(g_1, d_2)$ $dist(d_1, d_2)$
ГПС и ДПС	$dist(g_1, g_2)$	$dist(d_1, g_2)$; $dist(d_1, d_2)$	$dist(g_1, g_2)$; $dist(d_1, d_2)$

[a.] Расстояние между указанными точками; b. Центр ГПС в 1 последовательности; c. Центр ГПС во 2 последовательности; d. Центр ДПС в 1 последовательности; e. Центр ДПС во 2 последовательности

4. МНОЖЕСТВЕННОЕ ВЫРАВНИВАНИЕ

Выравнивание набора последовательностей происходит посредством последовательного попарного выравнивания различных последовательностей из набора. Перед этим набор сортируется в порядке уменьшения количества присутствующих в последовательностях обследований. Затем выполняется парное выравнивание двух первых последовательностей. Из всех вариантов их выравнивания выбираются первые N выравниваний в списке, отсортированном по увеличению штрафа. N – «память алгоритма», величина определяемая пользователем, увеличение которой улучшает итоговое множественное выравнивание, но ухудшает производительность алгоритма.

Затем к выравниванию добавляется следующая последовательность. Для этого для каждого из N выбранных выравниваний перебираются все возможные смещения новой последовательности и вычисляются соответствующие штрафы. Штраф для новой последовательности вычисляется как сумма штрафов её текущего положения относительно некоторых последовательностей, добавленных к выравниванию ранее. Среди всех положений новой последовательности во всех N существующих выравниваниях снова выбирается N лучших, то есть имеющих минимальный штраф. Так происходит до тех пор, пока не будут добавлены все последовательности. В результате работы алгоритма множественного выравнивания получаем N лучших выравниваний. В каждом выравнивании описаны целочисленные смещения каждой последовательности.

5. ТЕСТИРОВАНИЕ

Тестирование алгоритма проводилось на симулированных данных. Симуляция данных заключалась в определении параметров (математическое ожидание и стандартное отклонение) для каждой стадии. По этим параметрам формировался набор последовательностей, в которых длительность каждой стадии определялась как величина из нормального распределения с выбранными для этой стадии параметрами. Далее производилось сокрытие информации в каждой последовательности с сохранением только нескольких позиций, имитирующих моменты обследований и расположенных с определённым интервалом, длительность которого также соответствовала нормально распределённой величине с заданными параметрами.

На Рис. 2 приведён пример выравнивания 10 последовательностей процесса из 3 стадий (a, b, c). Математические ожидания продолжительностей стадий при симуляции были выбраны следующие (стандартное отклонение у всех стадий задано 1): a – 10, b – 3, c – 6. Моменты обследований выбирались со средним интервалом в 7 единиц при стандартном отклонении равном 3.

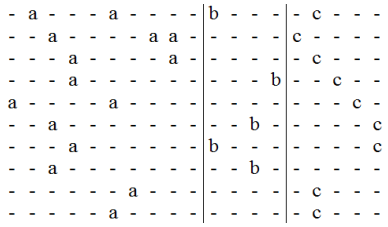

Рис. 2. Пример выравнивания 10 последовательностей

По результатам выравнивания можно оценить продолжительности стадий: a – от 9 до 10, b – от 4 до 5, c – около 5. Такие оценки в значительной мере соответствуют задаваемым исходным значениям при симуляции.

6. ЗАКЛЮЧЕНИЕ

Нами разработан подход к оценке продолжительности стадий заболевания по набору интервально-цензурированных данных. Планируется дальнейшая доработка описанного выше алгоритма, но уже сейчас можно заметить потенциал данного подхода к оценке длительности интервалов в неполных данных. Представленный подход применим к оценке временных показателей любых многостадийных процессов.

ЛИТЕРАТУРА

[1] Martino, A. Multivariate hidden markov models for disease progression / A. Martino, G. Guatteri, A.M. Paganoni // Statistical Analysis and Data Mining: The ASA Data Science Journal. – 2020. – Vol. 13(5). – P. 499-507.

[2] Tolles, J. Modeling epidemics with compartmental models / J. Tolles, T.B. Luong // Jama. – 2020. – Vol. 323(24). – P. 2515-2516.

[3] Андреев, Д.М. Стандартизация моделирования прогрессирования хронических заболеваний / Д.А. Андреев, Н.В. Хачанова, В.Н. Степанова, Е.Е. Башлакова, Е.П. Евдошенко, М.В. Давыдовская // Проблемы стандартизации в здравоохранении. – 2017. – № 9-10. – С. 12-23.

[4] Кондратьев, М.А. Методы прогнозирования и модели распространения заболеваний / М.А. Кондратьев // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 5. – С. 863-882.

[5] Маркович, Н.М. Оценка эпидемиологических показателей заболеваемости по косвенным данным / Н.М. Маркович, А.И. Михальский, В. Моргенштерн // Автоматика и телемеханика. – 1998. – № 6. – С. 153-162.

[6] Михальский, А.И. Моделирование заболеваний как обратная задача / А.И. Михальский // Труды 3-й Международной конференции «Высокие технологии, исследования, образование в физиологии, медицине и фармакологии». – СПб.: Издательство Политехнического университета, 2012. – Т. 2. – С. 213-216.

[7] Романюха, А.А. Математические модели в иммунологии и эпидемиологии инфекционных заболеваний / А.А. Романюха. – Москва: Бином. Лаборатория знаний, 2018. – 293 с.