УДК 621.57.087:004.616.1

ВЫБОР МЕТОДА ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ ДЛЯ ОЦЕНКИ СЕРДЕЧНО-СОСУДИСТОЙ ДЕЯТЕЛЬНОСТИ ПОДРОСТКОВ

А.И. Бых
Доктор физико-математических наук, профессор, заведующий кафедрой*
Е.В. Высоцкая
Кандидат технических наук, доцент*
Л.И. Рак
Кандидат медицинских наук, ведущий научный сотрудник ГУ «Институт охраны здоровья детей и подростков АМН Украины»
пр. 50-летия ВЛКСМ, 52-а, г. Харьков, 61153
Контактный тел.: (0572) 62-11-50
А.П. Порван
Научный сотрудник*
Е.Е. Болибок*
E-mail: volha_len@mail.ru
О.А. Сватенко*
*Кафедра биомедицинских электронных устройств и систем
Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, г. Харьков, 61166
Контактный тел.: (057) 702-13-64
E-mail: diagnost@kture.kharkov.ua

У статті надано порівняльну характеристику двом методам імпутації пропущених даних, необхідних для оцінки діяльності серцево-судинної системи та діагностики хронічної серцевої недостатності у дітей і підлітків з патологією міокарда

Ключові слова: імпутація даних, ЕМ-метод, метод лінійної регресії

В статье дана сравнительная характеристика двум методам импутации пропущенных данных, необходимых для оценки деятельности сердечной сосудистой системы и диагностики хронической сердечной недостаточности у детей и подростков с патологией миокарда

Ключевые слова: импутация данных, ЕМ-метод, метод линейной регрессии

In this article is given comparative description to two methods of the imputations skipped information necessary for the estimation of cardiovascular system and diagnostics chronic heart insufficiency on children and teenagers with myocardium pathology

Keywords: imputation of information, EM-method, linear regression method

1. Введение

Проблема хронической сердечной недостаточности (ХСН) является одним из приоритетных направлений развития современной кардиологии. Это связано как с неуклонным ростом данной патологии, так и с тяжелыми ее последствиями: сокращением продолжительности жизни и увеличением риска внезапных фатальных событий, инвалидизацией и ухудшением качества жизни больных [1].

В педиатрической практике немало заболеваний, возникающих на фоне как кардиальной, так и экстракардиальной патологии, которые могут привести к ухудшению работы сердца и развитию ХСН. Наиболее часто у детей и подростков ХСН формируется вследствие воспалительных заболеваний миокарда, вторичных кардиомиопатий и врожденных пороков развития [2].

Объективной проблемой на этапе обработки результатов является наличие пропусков данных у больных

в связи с различными возможностями исследований (в зависимости от сезона года, состояния аппаратуры и т.п.). Если игнорировать наличие пропусков в данных, то существенно вырастает риск получения неверных или незначимых результатов.

Проблему пропусков в данных исследователи решают по-разному. Некоторые исключают из рассмотрения наблюдения с пропущенными данными. Другие исследователи стремятся на этапе первичной обработки заполнить пропуски в уже имеющихся данных, для того чтобы восстановить исходную зависимость [3].

2. Анализ проблемы и постановка задачи исследования

В настоящее время для анализа медицинских данных, имеющих пропуски, широко используются узкоспециализированные инструментальные системы статистической обработки информации, такие как SPSS, STATISTICA, Systat, SAS, STADIA, Statgraphics, которые имеют большой набор статистических функций: факторный анализ, регрессионный анализ, кластерный анализ, многомерный анализ, критерии согласия и т.д. [4].

Можно выделить следующие группы методов импутирования (заполнения пропусков). К простым (неитеративным) алгоритмам, основанным на простых арифметических операциях, расстояниях между объектами, относятся: метод HotDeck, заполнение пропусков средним арифметическим, регрессионное моделирование пропусков и подбор в группе [5].

Самым простым методом является заполнение средним значением (модой, медианой или средним арифметическим), найденным по имеющимся данным. Он не требует применения специального программного обеспечения. Средние значения, вычисленные на исходном и преобразованном массивах, совпадают. Однако такого рода преобразование «усредняет» данные, уменьшая дисперсию признака, и, следовательно, показатели корреляции.

Метод ближайшего соседа предполагает, что пропуски будут заполнены различными значениями, полученными в результате оценивания расстояния между центроидами рассматриваемого набора данных. Такое заполнение еще называют заполнением с пристрастным выбором: считается, что объекты имеют одинаковые значения, если похожи по ряду прочих характеристик, отраженных в массиве. Недостатком данного метода является то, что он требует значительных вычислительных затрат. Вероятно, что предсказания будут неточными, если пропуск данных не имеет никакой закономерности.

При использовании метода многомерной регрессии строится модель линейной зависимости переменной, в которой необходимо заполнить пропуски, от ряда других имеющихся признаков. Регрессионные коэффициенты для каждого из предикторов находятся методом наименьших квадратов в массиве с полными данными. Подставляя значения предикторов в регрессионное уравнение, получают прогноз пропущенного показателя. Недостаток этого метода заключается в том, что в некоторых случаях могут быть пропущены не только значения переменной, которую нужно предсказать с помощью регрессии, но и значения предикторов – предсказание непосредственно на основе коэффициентов уравнения оказывается невозможным. Сложность применения данного метода еще и в том, что исследователь должен выбрать переменные, коррелирующие с рабочей переменной. В массиве может просто не оказаться достаточного количества предикторов. Предсказанные значения не содержат остатков, характерных для любых данных.

Сложные алгоритмы (итеративные) предполагают оптимизацию некоторого функционала, отражающего точность расчета подставляемых на место пропуска значений. Их делят на глобальные и локальные.

Особенностью локальных алгоритмов является то, что в оценивании (предсказании) каждого пропущенного значения участвуют полные наблюдения, находящиеся в некоторой окрестности предсказываемого объекта. К данной группе относятся алгоритмы Zet и Zet Braid.

Глобальные алгоритмы для оценивания каждого пропущенного значения оперируют всеми объектами рассматриваемой совокупности. К ним относятся метод Бартлетта, EM-оценивания (EM – expected value maximization) и Resampling (метод попарного сравнения) [6, 7].

Метод Бартлетта представляет собой алгоритм, включающий три итерации. На первой итерации пропуски заполняются некоторым начальным значением (например, средним арифметическим по имеющимся данным). На второй итерации для преобразованной переменной строится регрессионная модель. На заключительном этапе на основе полученного регрессионного уравнения предсказываются новые значения для пропусков.

Суть алгоритма Resampling заключается в том, что значения для пропусков выбираются из имеющихся случайным образом, с возвращением (когда значение может использоваться еще раз после выбора) или без него. После этого на всем массиве строится регрессионная модель, позволяющая предсказать значения для пробелов. Для всех предполагаемых предикторов находятся регрессионные коэффициенты и константа. Затем вычисляются итоговые значения регрессионных коэффициентов, по которым и будут предсказаны окончательные пропущенные значения.

В связи с этим целью настоящего исследования стало моделирование пропусков в базе данных по основным характеристикам сердечно-сосудистой деятельности у детей с патологией миокарда на основании методов импутирования пропущенных данных (линейной регрессии и EM-оценивания).

3. Экспериментальные исследования

Одной из задач эксперимента, наряду с оценкой потерь информации из-за неполноты данных, было сравнение качества импутирования данных с использованием EM-оценивания и регрессионного моделирования, так как при восстановлении данных с помощью сложных алгоритмов выборочные оценки чаще являются состоятельными и несмещенными.

Были рассмотрены морфофункциональные параметры сердца у 100 подростков 12-18 лет с патологией сердца воспалительного и невоспалительного генеза, 4% которых были пропущены.

Морфометрические характеристики сердца определялись по стандартной методике с помощью ультра-

звукового допплеровского исследования сердца на аппарате цифровой системы ультразвуковой диагностики SA-8000 Live (фирмы "Medison", Корея). Анализировались такие показатели: размеры диаметра корня аорты, левого предсердия (ЛП), правого желудочка (ПЖ), толщина миокарда (ТМ) задней стенки левого желудочка (ЛЖ), толщина межжелудочковой перегородки; конечный дистолический и конечный систолический размеры ЛЖ, конечный дистолический и конечный систолический объемы ЛЖ (КДО, КСО), фракция выброса ЛЖ (ФВ ЛЖ); ударный объем ЛЖ; минутный объем крови (МО); частота сердечных сокращений (ЧСС). Общее периферическое сопротивление сосудов (ОПСС) рассчитывалось по формуле

$$ОПСС = (Му \times 79{,}98) : МО,$$

где ОПСС – общее периферическое сопротивление сосудов, дин.с.см$^{-5}$;

Му – среднее артериальное давление, мм рт. ст.;

МО – минутный объем крови, л/мин.

Диастолическую функцию ЛЖ оценивали за такими временными и скоростными показателями: максимальная скорость раннего диастолического наполнения и предсердной систолы; время замедления раннего дистолического наполнения; время изоволюметрического расслабления. Оценивали также скорость кровотока через трикуспидальный, аортальный и легочный клапаны и градиент давления в магистральных сосудах по уравнению Бернулли.

Проводилась также стресс-эхокардиография с физической нагрузкой, после которой оценивались функциональные параметры ЛЖ.

Для получения идеального массива данных, не содержащего пропусков, изначально в анализ были включены только полные наблюдения. Регрессионная модель, построенная на этом массиве, в дальнейшем являлась эталонной для сравнения с ней моделей, полученных после заполнения пропусков двумя рассматриваемыми методами. Более эффективным можно считать тот метод импутирования, который обеспечит наибольшее приближение данных после импутирования и построенной на них регрессионной модели к идеальным показателям. Импутацию данных проводили с использованием пакета SPSS Statistics 17.0 [8].

Для того чтобы оценить потерю информации и качество импутирования из идеального массива размерностью 8×100 был создан отдельный массив с 4% пропусков для восьми переменных (табл. 1).

Таблица 1

Переменные, используемые в импутации данных

Переменная	Количество пропущенных значений
Возраст, полных лет	1
ЧСС, уд./мин.	1
Амплитуда Р-зубца, мкВ	1
ЛП, см	2
ПЖ, см	1
ОПСС, дин.с.см$^{-5}$	3
Скорость транстрикуспидального кровотока, см/с	5
ФВ ЛЖ (в пробе с физической нагрузкой), %	8

Пропущенные значения были импутированы двумя методами: методом EM-оценивания и методом линейного регрессионного моделирования, в результате чего было получено два массива полных наблюдений с искусственно вставленными значениями. Затем для каждого массива была отстроена линейная регрессионная модель.

Чтобы понять, какой из двух методов обеспечивает максимально точную импутацию, полученные модели сравнивались с эталонной (табл. 2).

В результате оценки потери информации, возникшей из-за исключения из регрессионного анализа неполных наблюдений, модель, построенная по импутированным данным, оказалась менее точной. Потери в точности во многом зависят от качества предсказания отсутствующих значений, которое можно будет оценить, так как истинные значения по всем характеристикам для каждого объекта нам известны.

Таблица 2

Исходные значения и их импутированные аналоги для сравниваемых методов

Переменная	Исходное значение	Значение, полученное метод EM-оценивания	Значение, полученное метод линейного регрессионного моделирования
Возраст, полных лет	15	16	18,5
ЧСС, уд./мин.	62	63	64
Амплитуда Р-зубца, мкВ	0,080	0,080	0,087
ЛП, см	2,65	2,70	2,68
	2,30	2,30	2,27
ПЖ, см	1,81	2,80	4,53
ОПСС, дин.с.см$^{-5}$	2443,80	2156,3	2142,72
	1126,20	800	762,51
	2615,70	2092,7	1975,23
Скорость транс-трикуспидального кровотока, см/с	97,80	90,4	89,24
	57,60	40,00	40,99
	84,60	90,4	80,84
	54,31	49,24	49,83
	76,50	76,5	77,82
ФВ ЛЖ (в пробе с физической нагрузкой), %	72,30	72,30	72,85
	71,00	72,90	66,11
	74,60	71,82	79,36
	64,00	74,00	63,88
	78,20	64,50	77,75
	69,60	77,50	73,58
	74,20	71,35	73,56
	78,23	72,6	77,82

Для определения качества импутирования был использован критерий Стьюдента для проверки равенства средних при уровне значимости p= 0,05 (табл. 3).

Таблица 3

Значения t-статистики двух сравниваемых методов импутации данных по переменным для матрицы 8х100

Сравниваемый метод	Переменная							
	Возраст, полных лет	ЧСС, уд./мин.	Амплитуда Р-зубца, мкВ	ЛП, см	ПЖ, см	ОПСС, кПа	Скорость транстрикуспидального кровотока, см/с	ФВ ЛЖ (в пробе с физической нагрузкой), %
Метод ЕМ-оценивания	0,035	0,005	0,000	0,014	0,212	0,148	0,119	0,049
Метод линейного регрессионного моделирования	0,123	0,011	0,065	0,000	0,517	0,170	0,158	0,026

Из рассчитанных значений видно, что средний показатель значимости массива, заполненного методом ЕМ-оценивания, выше (t=0,073), чем массива, заполненного методом линейного регрессионного моделирования (t=0,134). При этом уровень значимости полученных результатов больше 0,05 (p=0,394), что позволяет говорить об отсутствии значимой разницы между данными в идеальном массиве и данными, импутированными методом ЕМ-оценивания.

4. Выводы

Таким образом, для небольшой доли пропущенных значений при исследовании патологии сердечно-сосудистой системы у подростков рациональнее и эффективнее использовать метод ЕМ-оценивания. Рассчитанные значения t-статистики позволяют утверждать, что при 4% искусственно рассчитанных значениях переменных показатели отклоняются от истинных крайне незначительно, что позволяет ими пренебречь. Заполнение пропусков отдельных параметров с помощью математических методов позволяет с полным набором морфофункционального состояния сердца и других биохимических констант строить диагностическую и прогностическую модели ХСН. Применение ЕМ-метода позволяет не только восстановить пропущенные значения посредством двух этапного итеративного алгоритма, но и оценить средние значения для количественных переменных.

Литература

1. Гуревич, М.А. Хроническая сердечная недостаточность: руководство для врачей [Текст] / М.А.Гуревич. – М.: Практическая медицина, 2008. – 414 с.
2. Кржечковская, В. Заболевания сердечно-сосудистой системы детей и подростков [Текст] / В. Кржечковская, Р. Вахтангишвили. – М.: Феникс, 2006. – 508 с.
3. Злоба, Е. Статистические методы восстановления пропущенных данных [Текст] / Е. Злоба, И. Яцкив // Computer Modeling & New Technologies. – 2004. – Vol. 6. – P. 55–56 .
4. Rubin, D.B. Nested multiple imputation of NMES via partially incompatible MCMC [Текст] / D.B. Rubin // Statistica Neerlandica. – 2003. – Vol. 57. – P.3–18.
5. Чупеев, А.Н. Методы анализа значимости показателей при классификационном и прогностическом моделировании [Текст] / А.Н. Чупеев, О.Н. Чоповров, С.Ю. Брегеда // Вестник Воронежского государственного технического университета. – 2008. –Т. 4, №9. – С. 92–94 .
6. Крыштановский, А.О. Анализ социологических данных с помощью пакета SPSS [Текст] / А.О. Крыштановский. – М.: ГУ-ВШЭ, 2006. – 263 с.
7. Royston, P. Multiple imputation of missing values [Текст] / P. Royston // The Stata Journal. – 2004. – Vol. 4. – P. 227–241.
8. Programming and Data Management for SPSS Statistics 17.0 [Текст]: A Guide for SPSS Statistics and SAS Users. – Chicago: Raynald Levesque and SPSS Inc., 2007. – 576 p.