УДК 519.216.3:616.379-008.64

Некоторые математические подходы в построении моделей прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков

О.С. Кротова[1], А.И. Пиянзин[1,2], Л.А. Хворова[1], А.В. Жариков[1]

[1]Алтайский государственный университет (Барнаул, Россия)
[2]Алтайский государственный медицинский университет (Барнаул, Россия)

Some Mathematical Approaches to Develop Models for Prediction of Compensation and Decompensation Stages of Diabetes Mellitus among Children and Adolescents

O.S. Krotova[1], A.I. Piyanzin[1,2], L.A. Khvorova[1], A.V. Zharikov[1]

[1]Altai State University (Barnaul, Russia)
[2]Altai State Medical University (Barnaul, Russia)

В статье рассматривается задача прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков методами машинного обучения. Для проведения исследования разработано несколько математических моделей: логистическая регрессия, деревья решений и градиентный бустинг.

Информационное обеспечение моделей представлено «обезличенными» данными медицинского обследования детей и подростков Алтайского края, страдающих сахарным диабетом.

Выходным параметром моделей является стадия компенсации сахарного диабета, который может принимать значения: 0 — компенсация сахарного диабета, 1 — декомпенсация сахарного диабета. Задача прогнозирования стадии компенсации сахарного диабета у детей и подростков есть задача бинарной классификации.

В результате проведенного исследования сделано следующее: построены модели прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков на высокоуровневом языке программирования Python, подобраны оптимальные значения параметров для каждой модели, проведена оценка качества построенных моделей с помощью следующих метрик: точность, полнота, F-мера, чувствительность и специфичность.

Результаты данного исследования могут быть использованы специалистами для дополнительной диагностики детей и подростков Алтайского края, страдающих сахарным диабетом.

Ключевые слова: сахарный диабет, стадии компенсации и декомпенсации, методы классификации данных, моделирование.

The problem of prediction of compensation and decompensation stages of diabetes mellitus among children and adolescents using methods of machine learning is considered in the paper. There are several mathematical models used in the study: logistic regression, decision trees and gradient boosting.

The "de-identified" data of medical examination of children and adolescents of the Altai region suffering from diabetes mellitus are used to train the models in this study.

The output parameter of the models is the stage of diabetes mellitus compensation encoded with the following values: 0 — compensation of diabetes mellitus, 1 — decompensation of diabetes mellitus. This way, the prediction is the problem of binary classification.

The results of the conducted research are the following: models to predict the stages of compensation and decompensation of diabetes mellitus among children and adolescents are developed using the high-level Python programming language; optimal parameters are obtained for each model; prediction quality is estimated for each model using the following metrics: accuracy, completeness, F-measure, sensitivity, and specificity.

Professionals can use the obtained results for the supplementary diagnosis of children and adolescents of the Altay region who suffer from diabetes mellitus.

Key words: diabetes mellitus, stages of compensation and decompensation, data classification methods, modeling.

DOI 10.14258/izvasu(2018)4-15

Введение. Сахарный диабет является тяжелым хроническим заболеванием и с каждым годом все чаще встречается у детей и подростков.

Рост заболеваемости и высокая степень инвалидизации среди детей и подростков Алтайского края делают проблему всестороннего изучения, диагностирования и прогнозирования стадий сахарного диабета актуальной и практически значимой.

Сахарный диабет — системное гетерогенное заболевание, связанное с нарушением усвоения глюкозы и развивающееся вследствие абсолютного (1 тип) или относительного (2 тип) дефицита гормона поджелудочной железы — инсулина, который вначале вызывает нарушение углеводного обмена, а затем всех видов обмена веществ, что в конечном итоге приводит к поражению всех функциональных систем организма [1].

Состояние углеводного обмена определяется стадиями компенсации сахарного диабета — компенсацией и декомпенсацией. Компенсация сахарного диабета характеризуется близкими к нормальным показателями уровня глюкозы в крови. При декомпенсации сахарного диабета наблюдается повышенный уровень глюкозы в крови, который не поддается коррекции лекарственными препаратами.

В детском возрасте достаточно быстро наступает привыкание к гипергликемии — повышенному содержанию глюкозы в крови, что не вызывает заметного ухудшения самочувствия больного. Наличие различных осложнений, задержка физического развития и полового созревания являются поздними показателями длительной декомпенсации сахарного диабета [2]. Ранее выявление и прогнозирование стадий компенсации заболевания позволяют родителям и врачам проводить целенаправленные действия, помогающие сохранить здоровье ребенка и отсрочить инвалидизацию.

Актуальность и значимость проблемы определили цель исследования — построение моделей прогнозирования стадий компенсации и декомпенсации сахарного диабета у детей и подростков методами машинного обучения [3–6].

Данные для исследования размещены в информационной системе «Медицинская карта пациента», разработанной авторами [7], которая содержит «обезличенные» данные медицинского обследования детей и подростков Алтайского края, страдающих сахарным диабетом. Информационная система обеспечивает автоматическое формирование выборок данных пациентов по различным критериям.

Для построения моделей прогнозирования стадий компенсации и декомпенсации сахарного диабета была сформирована выборка данных, в которую вошли такие признаки, как рост, вес, температура, артериальное давление, частота сердечных сокращений, частота дыхания, стаж заболевания, показатели биохимического анализа крови. Результирующим параметром является стадия компенсации сахарного диабета, который на выходе модели может принимать значения: 0 — компенсация сахарного диабета, 1— декомпенсация сахарного диабета. Таким образом, задача прогнозирования стадии компенсации сахарного диабета у детей и подростков является задачей бинарной классификации.

Методы классификации данных. Для решения задач исследования было построено несколько математических моделей.

1. Логистическая регрессия. Логистическая регрессия используется для предсказания вероятности наступления некоторого события по значениям множества признаков.

Рассмотрим задачу бинарной классификации, где множество классов $Y = \{0,1\}$. Пусть p — вероятность некоторого события в бинарном случае. Отношение $\frac{p}{1-p}$ называется перевесом или преимуществом. Логарифм от отношения $\frac{p}{1-p}$ определяется как логит-функция вероятности:

$$logit(p) = \log\left(\frac{p}{1-p}\right).$$

Если $p(y = 1|x)$ — условная вероятность того, что отдельно взятый объект принадлежит классу 1 при наличии его признакового описания x, тогда

$$logit\left(p\left(y=1|x\right)\right) = \omega_0 x_0 + \omega_1 x_1 + \ldots + \omega_m x_m = \sum_{i=0}^{m} \omega_i x_i = \omega^T x,$$

где $\omega = (\omega_0, \ldots, \omega_m)$ — вектор весов.

Функция $\varphi(z)$, обратная к *logit*, называется логистической функцией или сигмоидой и позволяет предсказывать вероятность того, что определенный объект принадлежит отдельно взятому классу:

$$\varphi(z) = \frac{1}{1+e^{-z}},$$

где $z = \omega_0 x_0 + \omega_1 x_1 + \ldots + \omega_m x_m = \omega^T x$ — линейная комбинация весов и признаков объекта.

Вход сигмоидной функции интерпретируется как вероятность принадлежности отдельно взятого объекта i классу: $\varphi(z) = P(y = i|x, \omega)$.

Предсказанная вероятность конвертируется в бинарный результат:

$$\tilde{y} = \begin{cases} 1, \text{ если } \varphi(z) \geq 0{,}5, \\ 0, \text{ если } \varphi(z) \leq 0{,}5. \end{cases}$$

Для подбора параметров $\omega_0,\ldots,\omega_n$ используется метод максимального правдоподобия.

Зададим обучающую выборку, которая представляет собой набор пар $x^{(1)}, y^{(1)}\big),\ldots,\big(x^{(n)}, y^{(n)}\big)$, где $x^{(j)} \epsilon R^n$, $y^{(j)} \epsilon Y$.

Определим функцию правдоподобия L, которую необходимо максимизировать:

$$L(\omega) = P(y|x;\omega) = \prod_{j=1}^{n} P(y^{(j)} \mid x^{(j)};\omega) = \prod_{j=1}^{n} \left(\varphi\left(z^{(j)}\right)\right)^{y^{(j)}} (1-\varphi\left(z^{(j)}\right))^{1-y^{(j)}}.$$

Вследствие того, что значения функции правдоподобия $L(\omega)$ могут быть достаточно малыми, максимизируем не саму функцию $L(\omega)$, а ее логарифм, который также является функцией правдоподобия:

$$l(\omega) = \log L(\omega) = \sum_{j=1}^{n} [y^{(j)} \log(\varphi\left(z^{(j)}\right)) + \left(1-y^{(j)}\right) \log(1-\varphi\left(z^{(j)}\right))].$$

2. Деревья решений. Деревья решений подразумевают разбиение данных на классы путем принятия решений, основываясь на построении логических схем. Опираясь на признаки в «тренировочном» наборе данных, модель дерева решений обучается на иерархически организованной системе вопросов. При этом задаваемый вопрос на каждом последующем иерархическом уровне зависит от ответа, полученного на предыдущем уровне.

Начиная с корня дерева, данные расщепляются по признаку, который ведет к наибольшему приросту информации (information gain, IG). Процедура расщепления является итеративной и повторяется до достижения концевой вершины (листа). Объект будет относиться к определенному классу согласно метке, поставленной в соответствие данному листу.

Для того, чтобы расщепить узлы в самых информативных признаках, определим целевую функцию:

$$IG\left(D_p, f\right) = I\left(D_p\right) - \sum_{j=1}^{m} \frac{N_j}{N_p} I\left(D_j\right),$$

где f — признак, по которому выполняется расщепление, D_p — набор данных родительского узла, D_J — набор данных дочернего узла, I — критерий расщепления, N_p — общее число объектов в родительском узле, N_j — число объектов в дочернем узле, $j = 1,\ldots,m$.

Тогда возникает задача оптимизации, состоящая в максимизации прироста информации при каждом расщеплении:

$$IG\left(D_p, f\right) = I\left(D_p\right) - \sum_{j=1}^{m} \frac{N_j}{N_p} I\left(D_j\right) \to \max.$$

В бинарных деревьях решений обычно используются следующие критерии расщепления:

1) энтропия

$$I_H\left(t\right) = -\sum_{i=1}^{c} p(i \mid t) \log_2 p(i \mid t),$$

здесь $p(i \mid t)$ — доля объектов, принадлежащих классу i для отдельно взятого узла t;

2) мера неоднородности Джини — критерий, минимизирующий вероятность ошибочной классификации:

$$I_G\left(t\right) = \sum_{i=1}^{c} p(i \mid t)\left(1 - p\left(i|t\right)\right) = 1 - \sum_{i=1}^{c} p\left(i|t\right)^2;$$

3) ошибка классификации

$$I_E\left(t\right) = 1 - \max\left\{p\left(i|t\right)\right\}.$$

3. Градиентный бустинг. Еще одним подходом к решению задач классификации, рассмотренным авторами, является комбинирование моделей. В результате объединения нескольких классификаторов удается получить модель, обобщающая способность которой гораздо лучше, чем у каждого классификатора в отдельности.

Бустинг — это процесс последовательного построения классификаторов таким образом, что каждый последующий добавленный алгоритм, используя данные об ошибках, стремится компенсировать недостатки композиции всех предыдущих классификаторов. Модель градиентного бустинга строится в виде суммы деревьев решений:

$$f\left(x\right) = h_0 + \nu \sum_{j=1}^{M} h_j\left(x\right),$$

где h_0 — начальное приближение (константа), h_x — регрессионные деревья решений, $\nu \in \left(0;1\right]$ — параметр, регулирующий скорость обучения.

В процессе реализации алгоритма градиентного бустинга новые деревья добавляются путем минимизации эмпирического риска, заданного функцией потерь:

$$L\left(y, y'\right) = L\left(y, f\left(x\right)\right).$$

Для бинарной классификации функция потерь имеет вид:

$$L(y, y', y') = -\sum (y = k) \ln(\frac{\exp(y')}{\sum \exp(y')}).$$

Результаты моделирования. Построение моделей осуществлено на высокоуровневом языке программирования Python [8–10]. Каждый алгоритм классификации содержит несколько настраиваемых параметров оптимизации. В таблице 1 представлены оптимальные значения параметров моделей, которые подбирались «вручную» в результате многократного запуска и сравнения результатов работы моделей.

Таблица 1

Оптимальные значения параметров

Классификатор	Параметры оптимизации	Оптимальные значения
Логистическая регрессия	C	0.01
	tol	0.00001
	max_iter	1
Деревья решений	max_depth	[1; 100]
	max_features	sqrt
Градиентный бустинг	n_estimators	[50; 250]
	max_depth	[1; 100]
	max_features	sqrt

Сравнение и оценка качества работы построенных моделей осуществлялись с помощью таких метрик, как точность, полнота, F-мера, чувствительность и специфичность. В таблице 2 представлены значения метрик точности, полноты и F-меры для каждой построенной модели.

Таблица 2

Значения точности, полноты и F-меры построенных моделей

Модель	Метка класса	Точность	Полнота	F-мера
Логистическая регрессия	0	0.50	0.33	0.40
	1	**0.71**	**0.83**	**0.77**
	total	0.64	0.67	0.65
Деревья решений	0	0.67	0.22	0.33
	1	**0.71**	**0.94**	**0.81**
	total	0.69	0.70	0.65
Градиентный бустинг	0	1.0	0.11	0.20
	1	**0.69**	**1.0**	**0.82**
	total	0.79	0.70	0.61

Все модели показали высокие значения метрик для декомпенсации сахарного диабета (метка класса — 1). Значения этих же метрик для компенсации сахарного диабета (метка класса — 0) значительно ниже.

В медицинской статистике для анализа данных применяются такие метрики, как чувствительность и специфичность [3]. Метод исследования считается оптимальным, если он высоко специфичен и высоко чувствителен. Однако в реальности повышение чувствительности неизбежно сопровождается потерей специфичности и наоборот, повышение специфичности сопряжено со снижением чувствительности. Чувствительность (Se) определяется как способность диагностического метода давать правильный результат. Чувствительность является аналогом TPR (True Positive Rate) — доли истинно положительных объектов.

Специфичность (Sp) — это способность диагностического метода не давать при отсутствии заболевания ложноположительных результатов, которые определяются как доля истинно отрицательных результатов среди здоровых лиц в группе исследуемых. Аналогом специфичности считается FPR (False Positive Rate) — доля ложноположительных объектов.

Чувствительность и специфичность, как и другие используемые метрики, рассчитываются с помощью матрицы ошибок, представленной таблицей 3.

Таблица 3

Матрица ошибок

	$y=1$	$y=0$
$\hat{y}=1$	True Positive (TP)	False Positive (FP)
$\hat{y}=0$	False Negative (FN)	True Negative (TN)

При этом $Sp=\frac{TP}{TP+FN}$, $Se=\frac{FP}{FP+TN}$.

Результаты анализа построенных моделей на чувствительность и специфичность приведены в таблице 4.

Таблица 4

Оценка чувствительности и специфичности моделей

Модель	Чувствительность	Специфичность
Логистическая регрессия	**33%**	**71%**
Деревья решений	22%	71%
Градиентный бустинг	11%	69%

Заключение. Результаты исследования показали, что рассмотренные методы машинного обучения позволяют обнаружить «скрытые» закономерности и механизмы протекания сахарного диабета у детей и подростков и могут эффективно применяться для диагностики декомпенсации. Модели продемонстрировали высокие значения метрик для декомпенсации сахарного диабета и достаточно малые значения этих же метрик для компенсации сахарного диабета. Это может быть связано с тем, что выборка данных содержит недостаточное количество информации о компенсированных пациентах. Решением данной проблемы может быть [11–13]: 1) пополнение базы данных пациентов; 2) подключение дополнительных клинических и лабораторных показателей состояния больных; 3) разработка новых моделей прогнозирования.

Использование полученных в результате исследования моделей позволит в кратчайшие сроки определять стадии компенсации сахарного диабета, что улучшит процесс диагностики и лечения заболевания у детей и подростков на территории Алтайского края.

Библиографический список

1. Дедов И.И., Кураева Т.Л., Петеркова В.А., Щербачёва А.Н. Сахарный диабет у детей и подростков. — М., 2002.

2. Дедов И.И., Кураева Т.Л., Петеркова В.А. Инсулинотерапия сахарного диабета 1 типа у детей и подростков. — М., 2003.

3. Медик В.А., Токмачев В.С., Фишман Б.Б. Теоретическая статистика // Статистика в медицине и биологии. — М., 2002.

4. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. — М., 2015.

5. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. — М., 2013.

6. Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. — СПб., 2017.

7. Пиянзин А.И., Сидун Д.Ю., Назаркина О.М., Хворова Л.А., Малахова Т.И., Шарлаева Е.А., Левич К.А., Сапкина М.Р., Назаровская О.В. Информационные технологии в оценке липидного обмена у детей и подростков с сахарным диабетом 1 типа // Медицинский алфавит. — 2017.

8. Рашка С. Python и машинное обучение. — М., 2017.

9. Коэльо Л., Ричарт В. Построение систем машинного обучения на языке Python. — М., 2016.

10. Виндер П. Python для сложных задач: наука о данных и машинное обучение. — СПб., 2018.

11. Кротова О.С., Хворова Л.А. Применение нейронных сетей для диагностики заболевания сахарным диабетом детей и подростков на территории Алтайского края // МАК: Математики — Алтайскому краю : сборник трудов всерос. конф. по математике. — Барнаул, 2017.

12. Кротова О.С., Сидун Д.Ю. Современные компьютерные технологии в изучении сахарного диабета у детей и подростков // Молодежь — Барнаулу : материалы XVIII—XIX городской научно-практической конференции молодых ученых. — Ч. XIX. — Барнаул, 2018.

13. Концепция создания единой государственной информационной системы в сфере здравоохранения : приказ Минздравсоцразвития России от 28.04.2011 № 364 [Электронный ресурс]. — URL: http://www.consultant.ru/.