05.13.17

# ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ В ДИАГНОСТИКЕ САХАРНОГО ДИАБЕТА

**О.С. Кротова, А.И. Пиянзин, Л.А. Хворова**
Алтайский государственный университет, г. Барнаул

Сахарный диабет является одним из наиболее опасных хронических заболеваний, в патогенезе которого лежит недостаток инсулина в организме человека, вызывающий нарушение обмена веществ и патологические изменения в различных органах и тканях и, в итоге, приводящее к поражению всех функциональных систем организма. Распространенность сахарного диабета среди детей и подростков в России носит характер эпидемии: по данным Министерства здравоохранения РФ в 2015 году было зарегистрировано 37670 детей и подростков, к 2017 году их количество увеличилось до 43506. Цель исследования – построение моделей определения стадий компенсации и декомпенсации сахарного диабета. Предметом исследования являются медицинские данные детей и подростков Алтайского края, страдающих сахарным диабетом. Объект исследования – интеллектуальный анализ данных. Интеллектуальный анализ данных представляет собой совокупность методов классификации, прогнозирования и моделирования, основанных на алгоритмах машинного обучения. Нередко к интеллектуальному анализу данных относят статистические методы, применяемые для анализа большого объема информации. В медицинских исследованиях, как и в практической медицине, спектр решаемых задач очень широк, что позволяет использовать различные методы интеллектуального анализа данных. Использование построенных моделей в медицинских учреждениях ускорит процесс диагностики и лечения сахарного диабета у детей и подростков, избавит врача от долгой рутинной работы. Раннее выявление и прогнозирование стадий компенсации и декомпенсации заболевания позволят родителям и врачам проводить целенаправленные действия, позволяющие сохранить здоровье ребенка и отсрочить инвалидизацию.

*Ключевые слова: сахарный диабет, интеллектуальный анализ данных, моделирование, классификация.*

## ВВЕДЕНИЕ

При лечении сахарного диабета основное внимание уделяется состоянию углеводного обмена, которое определяется стадиями компенсации сахар-ного диабета – компенсацией и декомпенсацией.

Компенсация сахарного диабета характеризуется близкими к нормальным показателями уровня глюкозы в крови, хорошим самочувствием и минимальным риском возникновения осложнений. При декомпенсации сахарного диабета наблюдается высокий уровень глюкозы в крови, который не поддается коррекции лекарственными препаратами, в результате чего развиваются серьезные поражения органов и систем организма больного.

В детском возрасте быстро наступает привыкание к гипергликемии, что не вызывает заметного ухудшения самочувствия больного. Наличие различных осложнений, задержка физического и полового развития, являются поздними признаками длительной декомпенсации сахарного диабета. Целью лечения сахарного диабета является его компенсация [1].

Определение стадий компенсации для врача является достаточно долгим и рутинным процессом, поэтому целью исследования является построение моделей определения стадий компенсации и декомпенсации сахарного диабета у детей и подростков методами машинного обучения. Такие модели позволят врачу в кратчайшие сроки определять стадии компенсации и декомпенсации сахарного диабета, своевременно корректировать лечение.

## ОСНОВНАЯ ЧАСТЬ

Данные для исследования представлены в информационной системе «Медицинская карта пациента» [2], содержащей результаты медицинского обследования детей и подростков Алтайского края, страдающих сахарным диабетом. Для построения моделей использовались следующие признаки: рост, вес, температура, артериальное давление, частота сердечных сокращений, частота дыхания, показатели общего и биохимического анализа крови. Результи-рующим параметром является стадия сахарного диабета, который на выходе модели может принимать значения: 0 – компенсация сахарного диабета, 1 – декомпенсация сахарного диабета. Таким образом, задача определения стадий компенсации и декомпен-сации сахарного диабета является задачей бинарной классификации.

Данные медицинских исследований пациентов содержат много пропусков, что ведет к большой потере информации. Поэтому одна из задач иссле-дования – восстановление пропущенных значений в медицинских данных.

Для обработки и восстановления пропущенных значений выбран язык статистических вычислений и графики – R, в котором реализованы все необходимые функции для работы с пропущенными данными.

Анализ структуры пропущенных значений показал, что полные данные по всем признакам имеют только 97 пациентов из 153, представленных в информационной системе. Меньше всего пропу-щенных значений имеется среди клинических показателей,

максимальное количество пропущенных значений содержатся в переменной «СОЭ».

Существует множество способов восстановления данных. Для решения поставленной задачи нами выбран метод множественного восстановления пропущенных значений. Рассматриваемый метод реализован в пакете Mice языка программирования R.

В процессе исследования проведены три эксперимента, в каждом из которых случайным образом в полном наборе данных создавались 10 пропущенных значений в переменной «частота дыхания» и осуществлялось восстановление этих значений с помощью функций пакета Mice. В качестве статистического метода использовалась обобщенная линейная регрессия, построенная на признаках, имеющих наибольшую корреляцию с признаком «частота дыхания».

Полагая, что отклонение предсказанного значения от фактического в пределах нормы для возраста пациента допустимо, ошибка восстановления соста-вила 13%. В 10% случаях фактическое значение параметра «частота дыхания» отличается от нормы, что требует привлечения дополнительной информа-ции о пациенте для оценки результата восстановления значения признака.

Восстановленные и обработанные данные позволят снизить процент потери информации и улучшить точность определения стадий компенсации и декомпенсации сахарного диабета у детей и подростков Алтайского края.

Для построения моделей определения стадий компенсации и декомпенсации сахарного диабета использовались методы ансамблевого обучения – адаптивный бустинг и бэггинг. Особенность ансамб-левых методов состоит в том, что они объединяют в себе несколько классификаторов, что способствует увеличению точности классификации данных.

**Адаптивный бустинг.** Модель адаптивного бустинга строится на нескольких простых классификаторах и использует всю обучающую выборку без разбиений. На каждом шаге объектам, которые были классифицированы неправильно, присваивается больший вес, а объектам, классифицированным верно, назначается меньший вес. Таким образом, каждый следующий классификатор сфокусирован на объектах, которые ранее были классифицированы неправильно, т.е. обучается на ошибках предыдущего классификатора.

Алгоритм, лежащий в основе адаптивного бустинга можно рассмотреть в виде псевдокода [3]:

Шаг 1. Элементам весового вектора $\omega$ присваиваются равномерные веса:

$$\sum_i \omega_i,\ i = 1, \dots, m.$$

Шаг 2. Осуществляется обучение простого классификатора:

$$C_j = \text{train}(X, y, x),$$

где $X$ – множество объектов, $y$ – значение выхода, $x$ – вектор признакового описания объекта, $j$ – номер объекта, $j = 1, .., m$.

Шаг 3. Определяются метки классов:

$$\hat{y} = \text{predict}(C_j, X).$$

Шаг 4. Вычисляется взвешенная частота появления ошибок – скалярное произведение вектора весов на вектор ошибок: $\varepsilon = \omega \cdot (\hat{y} \neq y)$. Вектор $(\hat{y} \neq y)$ состоит из 0 и 1: 1 означает, что прогноз ошибочный, 0 назначается в случае, когда прогноз верный.

Шаг 5. Вычисляется коэффициент $b_j$:

$$b_j = 0.5\, log\, \frac{1-\varepsilon}{\varepsilon}.$$

Шаг 6. Определяются весовые коэффициенты: $\omega := \omega \times exp(-b_j \times \hat{y} \times y)$, $(\times)$ – поэлементное умножение.

Шаг 7. Нормализуется вектор весов:

$$\omega := \frac{-\omega}{\sum_i \omega_i}$$

Шаг 8. Вычисляется итоговый прогноз:

$$\hat{y} = \sum_{j=1}^{m} (b_j \times predict(C_j, X)) > 0.$$

**Бэггинг.** Алгоритм бэггинг-классификации можно описать тремя этапами.

На первом этапе осуществляется разбиение исходных данных на подмножества. Пусть множество объектов $X$ состоит из $N$ элементов. Выберем $N$ раз произвольный объект из множества $X$. Причем, каждый раз мы выбираем объект из $N$ исходных объектов. Поскольку разбиение осуществляется случайным образом, то наборы объектов в подмно-жествах всегда будут разными: некоторые объекты попадут в несколько подмножеств, а некоторые не попадут ни в одно. Получившиеся в результате разбиения подмножества объектов называются бутстрап-выборками.

Предположим, что в результате разбиения исходных данных мы имеем $X_1, \dots, X_m$ подмножеств. На втором этапе для каждого подмножества строится классификатор $a_i$, $i = 1, \dots, m$.

Третий этап заключается в построении итогового классификатора α, выход которого будет равен среднему значению выходов классификаторов $a_i$, $i = 1, \dots, m$:

$$\alpha = \frac{1}{m} \sum_{i=1}^{m} a_{i.}$$

Построение моделей адаптивного бустинга и бэггинг-классификатора осуществлялось на языке программирования Python с применением классов библиотеки Scikit-learn. Разбиение исходных данных на обучающую и тесовую выборки выполнялось в процентном соотношении 70:30.

В качестве базового классификатора использовалось дерево решений, глубина которого равна 10. В рассматриваемой задаче, оптимальным для построе-

ния бэггинг-модели является ансамбль из 500 деревьев решений, для модели адаптивного бустинга – 700 деревьев.

Сравнение и оценка качества моделей проведены с помощью таких метрик как точность, полнота и *F*-мера.

Значение точности, полноты и *F*-меры для построенных моделей приведены в таблице 1 (метка класса 0 – компенсация сахарного диабета, 1 – декомпенсация сахарного диабета).

**Табл.1. Значения метрик точности, полноты и *F*-меры**

| Модель | Метка класса | Точность | Полнота | *F*-мера |
|---|---|---|---|---|
| Адаптивный бустинг | 0 | 0.80 | 0.44 | 0.57 |
| | 1 | 0.77 | 0.94 | 0.85 |
| | total | 0.78 | 0.78 | 0.76 |
| Бэггинг | 0 | **1.00** | **0.44** | **0.62** |
| | 1 | **0.78** | **1.00** | **0.88** |
| | total | **0.86** | **0.81** | **0.79** |

Результаты исследования показали, что значения метрик для ансамблевых методов выше, чем для алгоритмов, использовавшихся ранее в работах [4–5]. Наилучший результат показал бэггинг (табл. 1).

## ЗАКЛЮЧЕНИЕ

Применение построенных моделей позволит ускорить процесс диагностики и лечения сахарного диабета у детей и подростков. В дальнейшем плани-руется исследовать другие состояния заболевания, разработанные программные модули внедрить в медицинские учреждения соответствующего про-филя, и создать мобильное приложение, которое позволит вести диалог «ребенок-родители-врач».

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Дедов, И.И. Сахарный диабет в Российской Федерации: проблемы и пути решения [Текст] / И.И. Дедов. – М.: Эндокринологический научный центр РАМН, 1998. – 12 с.

2. Кротова, О.С. Современные компьютерные технологии в изучении сахарного диабета у детей и подростков [Текст] / Д.Ю. Сидун // Молодежь – Барнаулу: материалы XVIII – XIX городской научно-практической конференции молодых ученых. 2018. – С. 385–387.

3. Рашка, С. Python и машинное обучение [Текст] / С. Рашка. – М.: ДМК Пресс, 2017. – 418 с.

4. Кротова, О.С. Применение машинного обучения в изучении сахарного диабета у детей и подростков [Текст] / О.С. Кротова // Материалы 56-й Международной научной студенческой конференции МНСК-2018. – 2018. – С. 238.

5. Кротова, О.С. Методы и подходы глубокого обучения в изучении сахарного диабета у детей и подростков [Текст] / А.И. Пиянзин, Л.А. Хворова // Сборник трудов всерос. конф. по математике: МАК: Математики – Алтайскому краю. – 2018. – С. 327–328.

*Кротова Ольга Сергеевна – магистрант 1 курса факультета математики и информационных технологий, Алтайский государственный университет, тел. 8(923)7902542, e-mail: kr.olga0910@gmail.com.*

*Пиянзин Алексей Илларионович – кандидат медицинских наук, доцент кафедры информатики, Алтайский государственный университет, тел. 8(903)9481297, e-mail: bio7777777@mail.ru.*

*Хворова Любовь Анатольевна – кандидат технических наук, доцент, заведующий кафедрой теоретической кибернетики и прикладной математики, Алтайский государственный университет, тел. 8(913)2325206, e-mail: khvorovala@gmail.com.*

# DATA MINING IN THE DIAGNOSTICS OF DIABETES

**O.S. Krotova, A.I. Piyanzin, L.A. Khvorova**
*Altai State University, Barnaul*

Abstract – The article is devoted to the problems of diabetes. The purpose of the study is to build models for classify the stages of compensation and decompensation of diabetes by methods of data mining. The subject of the study are medical data of children of the Altai Krai who have diabetes. Using the constructed models will improve the process of diagnosis and treatment of diabetes in children.

Index terms: diabetes, data mining, modeling, classification.

## REFERENCES

1.Dedov, I.I., *Diabetes Mellitus in the Russian Federation: Problems and Solutions*. Moscow: Endocrinology Center of the Russian Academy of Medical Science, 1998.

2. Krotova, O.S., and D.Yu. Sidun"Modern computer technologies in the study of diabetes in children and adolescents," *19h City Scientific-Conference: Youth – Barnaul*, Barnaul, Russia, pp. 385–387, 2017.

3. Raschka, S., *Python Machine Learning*, 1st ed. Birmingham: Packt Publishing, 2015.

4. Krotova, O.S., "Application of machine learning in the study of diabetes in children and adolescents," *56th International Scientific Student Conference ISCC-2018*, Novosibirsk, Russia, p. 238, Apr. 2018.

5. Krotova, O.S., A.I. Pianzin, and L.A. Khvorova, "Methods and approaches of in-depth learning in the study of diabetes in children and adolescents," *Russian Conference on Mathematics: MAK: Mathematicians – Altai Krai*, Barnaul, Russia, pp. 327–328, Jun. 2018.

*Krotova Olga Sergeevna– masters's degree student of the faculty of mathematics and information technologies, Altai State University, tel. 8(923)7902542, e-mail: kr.olga0910@gmail.com.*

*Piyazin Alexey Illarionovich–candidate of medical sciences, associate professor of thechair informatics, Altai State University, tel. 8(903)9481297, e-mail: bio7777777@mail.ru.*

*KhvorovaLyubovAnatolievna–candidate of technical sciences, associate professor, head of the chair TCAM, Altai State University, tel. 8(913)2325206, e-mail: khvorovala@gmail.com*