УДК 519.8:616.379

Математические модели и алгоритмы машинного обучения в диагностике осложнений сахарного диабета первого типа

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

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

Mathematical Models and Machine Learning Algorithms in the Diagnosis of Complications of Type 1 Diabetes Mellitus

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

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

В статье рассматривается проблема диагностики диабетической полинейропатии — одного из наиболее ранних и опасных осложнений сахарного диабета у детей и подростков. Целью исследования является разработка моделей диагностики диабетической полинейропатии у детей и подростков на основе различных медицинских данных. Модели позволят диагностировать осложнение без применения нейрофизиологических методов исследования, что даст возможность использовать их в фельдшерско-акушерских пунктах в сельской местности и применять в качестве системы поддержки принятия врачебных решений. В ходе исследования проведен обзор и анализ научных публикаций отечественных и зарубежных ученых по теме проводимого исследования, обработан большой набор текстовых медицинских данных. Создана база данных, осуществлен анализ признаков и построена модель, определяющая наличие диабетической полинейропатии у детей и подростков, страдающих сахарным диабетом 1 типа. Достигнутая точность качества классификации позволяет утверждать, что методы машинного обучения могут применяться для поиска скрытых зависимостей в развитии и течении осложнений сахарного диабета.

Ключевые слова: математическое моделирование, машинное обучение, анализ данных, классификация, сахарный диабет.

The paper deals with the problem of diabetic polyneuropathy diagnosing. This is one of the earliest and most dangerous complications of diabetes among children and adolescents. The research aims to develop models for diagnosing diabetic polyneuropathy in children and adolescents based on various medical data. The developed models will make it possible to diagnose a complication without using neurophysiological research methods. Therefore, the proposed models can be used in small medical and obstetrical stations in rural areas as well as a support system for making medical decisions. In the course of the study, a review and analysis of scientific publications of domestic and foreign scientists on the topic of the research are carried out. A large set of textual medical data is processed, then a database is created, features are analyzed, and a model is developed to reveal the presence of diabetic polyneuropathy in children and adolescents with type 1 diabetes mellitus. The achieved quality of the classification model allows us to assert that machine learning methods can be used to find hidden dependencies in the development and course of complications of diabetes mellitus.

Key words: mathematical modeling, machine learning, data analysis, classification, diabetes mellitus.

DOI: 10.14258/izvasu(2021)1-16

Введение. Сахарный диабет — опасное хроническое заболевание, которое без регулярного контроля, ведения здорового образа жизни и правильного лечения приобретает тяжелое течение, сопровождающееся высоким риском возникновения серьезных осложнений.

Одним из наиболее распространенных микрососудистых осложнений сахарного диабета 1 типа у детей и подростков является диабетическая полинейропатия. Диабетическая полинейропатия характеризуется комплексом клинических и субклинических синдромов, вызванных поражением периферических

нервных волокон, приводит к ухудшению качества жизни и нередко становится причиной ранней инвалидности [1].

Скрининг осложнений сахарного диабета у детей и подростков проводится во время прохождения стационарного лечения. Основными методами диагностики диабетической полинейропатии являются опрос и осмотр больного. С целью получения более объективной информации проводится электронейромиография, которая позволяет выявить признаки поражения двигательных и чувствительных нервных волокон верхних и нижних конечностей [2].

Цель исследования — реализация математических моделей диагностики диабетической полинейропатии у детей и подростков на основе различных медицинских данных. Такие модели позволят диагностировать осложнения без применения нейрофизиологических методов исследования, что даст возможность использовать их в фельдшерско-акушерских пунктах в сельской местности для ранней диагностики диабетической полинейропатии у детей и подростков. Разрабатываемые модели могут применяться и в качестве системы поддержки принятия врачебных решений в диагностически неясных случаях.

Актуальность и практическую значимость проводимого исследования определяют: 1) быстрый рост заболеваемости и ранняя инвалидизация; 2) необходимость реализации персонализированного подхода к ранней диагностике осложнений сахарного диабета; 3) потребность в применении средств интеллектуального анализа больших данных для глубокого изучения структуры заболевания и его патогенеза.

Для программной реализации всех этапов исследования выбран язык программирования Python.

Анализ современного состояния исследований в области применения различных математических методов и подходов в диагностике и изучении диабетической полинейропатии показал, что большинство современных исследований, связанных с диабетической полинейропатией, не являются междисциплинарными, в них не используются современные математические методы и информационные технологии.

Результатом поиска публикаций, посвященных выявлению факторов риска возникновения диабетической полинейропатии у детей и подростков, страдающих сахарным диабетом 1 типа, стали два исследования. Первое исследование проведено учеными-медиками из Смоленской области [3]. В результате исследования учеными найдены статистически значимые отличия в значениях лабораторных и клинических показателей у детей с полинейропатией и без нее. Второе исследование было проведено турецкими учеными: с помощью линейного регрессионного анализа ими установлена связь между наличием диабетической полинейропатии и диабетического кетоацидоза [4].

В работе [5] ученые из Дании, применяя методы интеллектуального анализа данных, показали, что есть существенные отличия в значениях лабораторных показателей взрослых людей с разными осложнениями сахарного диабета 1 типа, но возникновение одного осложнения никак не связано с возникновением другого. Большое количество зарубежных публикаций посвящено изучениям полинейропатии у взрослых людей, страдающих сахарным диабетом 2 типа. В работе [6] ученые из Китая описывают результаты применения модели логистической регрессии для выявления факторов, влияющих на развитие полинейропатии у жителей городов Ухань и Чаншу. Аналогичные исследования ведутся по прогнозированию диабетической полинейропатии у пациентов с сахарным диабетом 2 типа. Так, например, в работе [7] предложена модель, предсказывающая тяжесть заболевания на основе пола, возраста пациента, стажа заболевания, дозировки принимаемых лекарственных препаратов, уровня основного метаболита витамина D в крови и гликированного гемоглобина.

Анализ текстовой медицинской информации. Данные для проведения исследования представлены в виде 3204 обезличенных медицинских выписок из историй болезни детей и подростков с сахарным диабетом 1 типа. Выписка из истории болезни представляет собой электронный документ, составленный в текстовом редакторе Microsoft Word. Часть данных, например результаты анализов, содержится в табличной форме, другая часть — в форме записей врача на естественном языке. Важнейшим этапом исследования является структурирование информации из медицинских выписок и создание базы данных.

Для формирования базы данных из медицинских выписок необходимо решить несколько задач: извлечение числовых и строковых данных из таблиц, нахождение в тексте упоминаний медицинских концептов (диагноз, сопутствующие заболевания, жалобы и симптомы и др.), извлечение числовых характеристик из текста.

Для решения поставленных задач выбран язык программирования Python (в частности, модуль re предназначен для работы с регулярными выражениями; библиотеки: docx — для обработки файлов, созданных в текстовом редакторе Microsoft Word и NLTK — для обработки естественного языка).

Важным признаком для данного исследования является наличие у пациента характерных для полинейропатии симптомов. Основной причиной диабетической полинейропатии считается хроническая гипергликемия. Основными проявлениями диабетической полинейропатии являются наличие болевого симптома, парестезии и уменьшение сухожильных рефлексов [1].

С помощью библиотеки NLTK языка программирования Python проведен анализ жалоб пациен-

тов на наличие гипергликемии, болевых ощущений и других симптомов, характерных для полинейропатии, в двух выборках (выборка 1 — выписки, в которых зафиксировано наличие полинейропатии, выборка 2 — выписки, в которых не зафиксировано наличие полинейропатии) [8]. Результаты анализа представлены в таблице 1.

Таблица 1

Анализ распространенности симптомов у пациентов

Симптом	Выборка 1	Выборка 2
Периодическая/стойкая гипергликемия	653 (59,4 %)	736 (35 %)
Болевой синдром (боли в конечностях, головные боли)	621 (56,4 %)	569 (27 %)
Другие симптомы (судороги в мышцах, слабость в конечностях, онемения, покалывания и др.)	606 (56,9 %)	642 (30,5 %)

Рассматриваемые симптомы значительно чаще встречаются у пациентов с диабетической полинейропатией, чем у пациентов, не имеющих осложнения. Однако каждый из симптомов сопутствует полинейропатии только в половине случаев.

Методы и модели. База данных содержит большое количество различных характеристик пациентов: возраст, жалобы, длительность заболевания, показатели общего анализа крови, общего анализа мочи, биохимического анализа крови и др. Однако не все имеющиеся признаки важны для построения моделей. Одной из задач этапа подготовки данных к моделированию является удаление малоинформативных, шумовых признаков.

Алгоритмы отбора признаков используются для автоматического отбора подмножества наиболее релевантных признаков. Последовательные алгоритмы отбора признаков относятся к семейству жадных алгоритмов поиска, которые используются для сведения начального d-мерного пространства признаков к k-мерному подпространству признаков, где $k<d$.

Для реализации отбора признаков выбран метод последовательного обратного отбора. Алгоритм, используя заданную модель классификации и критерий J, последовательно удаляет признаки из признакового множества до тех пор, пока не останется заданное количество признаков. Критерием является разность между значениями метрики качества классификации после и до удаления отдельно взятого признака [9]. Алгоритм последовательного обратного отбора можно выразить в трех шагах:

1. Инициализировать алгоритм при $k=d$, где d — размерность исходного признакового пространства.

2. Определить признак $\overline{x}$, который максимизирует заданный критерий: $\overline{x} = \arg\max J(X_k - x)$, $x \in X_k$.

3. Удалить признак $\overline{x}$ из набора признаков: $X_k = X_k - \overline{x}$, $k = k - 1$.

Алгоритм заканчивает работу, если k равняется числу требуемых признаков, иначе возвращается к шагу 2.

В качестве классификаторов выбраны: нелинейный метод опорных векторов (SVM) с ядром из функции радиального базиса и метод k Ближайших соседей из библиотеки Scikit-learn языка Python. В качестве критерия отбора выбрана доля правильных ответов. Предварительно проведены очистка и стандартизация данных.

На рисунках 1 и 2 приведены графики, отражающие зависимость качества классификации от количества используемых для обучения моделей признаков.

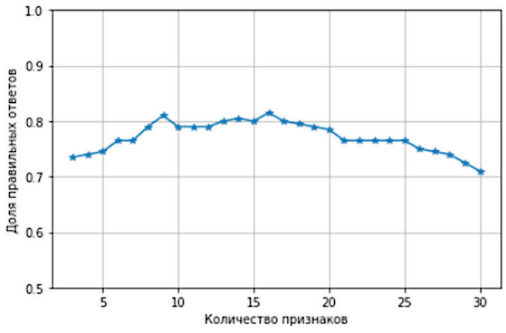

Рис. 1. Зависимость качества классификации от количества признаков, используемых для обучения модели SVM

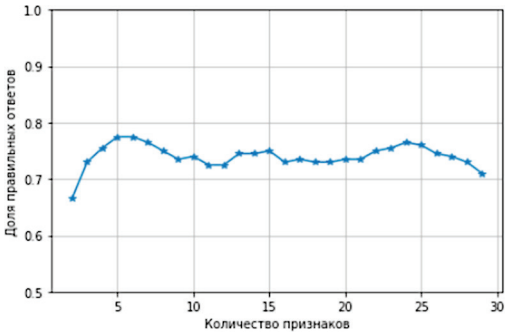

Рис. 2. Зависимость качества классификации от количества признаков, используемых для обучения модели k Ближайших соседей

В случае использования в качестве классификатора нелинейного метода опорных векторов максимальная доля правильных ответов составила приблизительно 81 % при обучении модели на 9 или 16 признаках. Для модели *k* Ближайших соседей максимальная доля правильных ответов составила 79 %.

Таким образом, определены наиболее значимые признаки: возраст пациента, длительность заболевания, наличие периодических или стойких гипергликемий, болевого синдрома, других симптомов, которые могут быть вызваны диабетической полинейропатией, уровень гликированного гемоглобина, содержание лимфоцитов, холестерина, щелочной фосфатазы, калия, сегментоядерных нейтрофилов, липопротеинов низкой плотности, билирубина, триглицеридов, калия и эритроцитов в крови. Выборка полных данных по отобранным признакам содержит 1993 наблюдения, в 751 из которых зафиксировано наличие диабетической полинейропатии.

Метод опорных векторов (supportvectormachine, SVM) — широко используемый и мощный алгоритм машинного обучения. SVM хорошо подходит для классификации сложных, но небольших по объему наборов данных.

Основная идея метода заключается в поиске границы решения — гиперплоскости, разделяющей объекты на два множества [10]. Две параллельные гиперплоскости строятся по обе стороны от границы решения. Алгоритм работает в предположении, что чем больше расстояние между гиперплоскостями, тем меньше будет средняя ошибка классификатора, поэтому метод также называют методом классификации с максимальным зазором.

Граница решения задается линейным уравнением $\omega^T x = 0$, где ω — вектор весовых коэффициентов. Положительную и отрицательную гиперплоскости, которые параллельны границе решения, можно задать уравнениями:

$$\omega_0 + \omega^T x_{pos} = 1, \quad (1)$$

$$\omega_0 + \omega^T x_{neg} = -1. \quad (2)$$

Зазором называют расстояние между положительной и отрицательной гиперплоскостями. Вычитая уравнение (1) из уравнения (2), получим:

$$\omega^T (x_{pos} - x_{neg}) = 2. \quad (3)$$

Выполним нормализацию по величине вектора весов ω, полагая, что его норма определяется следующим образом: $||\omega|| = \sqrt{\sum_{i=1}^{m} \omega_j^2}$.

Разделив выражение (3) на норму вектора ω, получим следующее соотношение: $\frac{\omega^T (x_{pos} - x_{neg})}{||\omega||} = \frac{2}{||\omega||}$.

Левую часть уравнения интерпретируют как расстояние между положительными и отрицательными гиперплоскостями и именуют как зазор.

Таким образом, задача сводится к максимизации зазора путем максимизации правой части уравнения с учетом предположения, что все объекты классифицированы правильно.

$$\omega_0 + \omega^T x^{(i)} \geq 1, \text{ если } y^{(i)} = 1, \quad (4)$$

$$\omega_0 + \omega^T x^{(i)} \leq -1, \text{ если } y^{(i)} = -1. \quad (5)$$

Из неравенств (4) и (5) следует, что все отрицательные образцы должны попасть на одну сторону от отрицательной гиперплоскости, а все положительные — остаться за положительной гиперплоскостью.

На практике проще минимизировать обратный член $\frac{1}{2}||\omega||^2$ с помощью методов квадратичного программирования.

Посредством модификации алгоритма — добавления ядра с помощью SVM можно решать нелинейные задачи. Ключевая идея, лежащая в основе ядерных методов для решения нелинейных задач, заключается в том, чтобы создать нелинейные комбинации исходных линейно неразделимых признаков и с помощью функции отображения φ спроецировать их на пространство более высокой размерности, где они становятся линейно разделимыми.

Сложность подхода на основе функции отображения состоит в том, что конструирование новых признаков в вычислительном плане очень затратно, поэтому на практике скалярное произведение $((x^i)^T, x^j)$ заменяется функцией ядра. Такой метод замены называется ядерным трюком: $K((x^i)^T, x^j) = (\varphi(x^i)^T, \varphi(x^j))$.

Для проведения исследования выбран SVM с ядром RBF (радиальная базисная функция), также известное как гауссовское ядро:

$$K_{rbf}(x^i, x^j) = \exp\left(-\frac{(x^i - x^j)^2}{2\sigma^2}\right) = \exp(-\gamma ||x^i - x^j||^2),$$

где $\gamma = \frac{1}{2\sigma^2}$ — свободный параметр — гиперпараметр, оптимальные значения которого подбираются под каждую отдельную задачу.

Для нахождения оптимальных значений гиперпараметров модели использовался поиск по сетке. Оптимизационный поиск по сетке, или сеточный поиск — это исчерпывающий поиск в заданном вручную подмножестве гиперпараметрического пространства с целью нахождения оптимальной комбинации ги-

перпараметров. Алгоритм поиска по сетке реализован в классе GridSeachCV() библиотеки Scikit-learn.

Результаты исследования. Оценка качества классификации проведена с использованием таких метрик, как точность, полнота, *F*-мера, чувствительность и специфичность. В тестовую выборку данных вошли 598 наблюдений: 384 наблюдения из класса -1 и 214 из класса 1. Результаты тестирования модели на данных, не используемых при обучении, приведены в таблице 2.

Таблица 2

Значения метрик качества классификации для полученной модели

Метка класса	Точность	Полнота	F-мера	Чувствительность	Специфичность
-1	0.76	0.82	0.79	0.62	0.82
1	0.69	0.62	0.65		

Несмотря на то что модель демонстрирует хорошее качество классификации на новых данных, достигнутых результатов недостаточно, для того чтобы использовать полученную модель во врачебной практике. Однако полученные значения метрик качества классификации показывают, что с помощью методов машинного обучения можно проводить диагностику диабетической полинейропатии без применения нейрофизиологических методов исследования.

Заключение. Проведены обзор и анализ отечественных и зарубежных научных публикаций по теме проводимого исследования, обработан большой набор текстовых медицинских данных, создана база данных, осуществлен анализ признаков и построена модель, определяющая наличие диабетической полинейропатии у детей и подростков, страдающих сахарным диабетом 1 типа.

Достигнутая точность качества классификации позволяет утверждать, что методы машинного обучения могут применяться для поиска скрытых зависимостей в развитии и течении осложнений сахарного диабета.

Библиографический список

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

2. Алимова И.Л. Диабетическая нейропатия у детей и подростков: нерешенные проблемы и новые возможности // Российский вестник перинатологии и педиатрии. 2016. № 3. DOI: 10.21508/1027-4065-2016-61-3-114-123.

3. Алимова И.Л., Лабузова Ю.В. Диабетическая полинейропатия у детей и подростков // Российский вестник перинатологии и педиатрии. 2009. № 6.

4. Turkyilmaz H., Guzel O., Edizer S., Unalp A. Evaluation of polyneuropathy and associated risk factors in children with type 1 diabetes mellitus // Turk J. MedSci., 2017. Vol. 47. DOI: 10.3906/sag-1601-183.

5. Bjerg L., Hulman A., Charles M., Jorgensen M.E., Witte D.R. Clustering of microvascular complications in Type 1 diabetes mellitus // J. Diabetes Complications. 2018. Vol. 32. DOI: 10.1016/j.jdiacomp.2018.01.011.

6. Qin L., Niu J.Y., Zhou J.Y., Zhang Q.J. et al. Prevalence and risk factors of diabetic peripheral neuropathy in Chinese communities // Zhonghua Liu Xing Bing XueZaZhi. 2019. Vol. 40. DOI: 10.3760/cma.j.issn.0254-6450.2019.12.014. (in Chinese).

7. Fitri A., Sjahrir H., Bachtiar A., Ichwan M., Fitri F.I., Rambe A.S. Predictive Model of Diabetic Polyneuropathy Severity Based on Vitamin D Level // Open Access Maced J. Med Sci. 2019. Vol. 7(16). DOI: 10.3889/oamjms.2019.454.

8. Krotova O.S., Moskalev I.V., Nazarkina O.M., Khvorova L.A. Diagnostics of Diabetic Polyneuropathy in Children and Adolescents Using Data Mining Methods // Journal of Physics: Conference Series. 2020. Vol. 1615. DOI:10.1088/1742-6596/1615/1/012015.

9. Рашка С. Python и машинное обучение. М., 2017.

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