УДК 654.9:615.478

КЛАССИФИКАЦИЯ ДАННЫХ В СИСТЕМЕ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ ПРОФИЛАКТИКИ САХАРНОГО ДИАБЕТА 2-ГО ТИПА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Л.К. Андреасян

Национальный политехнический университет Армении

Согласно статистике Всемирной организации здравоохранения, более трети взрослого населения имеют симптомы развития болезни сахарного диабета 2-го типа (СД2Т). Научные достижения в информационно-телекоммуникационных технологиях, развитие технологий IoT (Intertnet of Things) и широкое применение методов искусственного интеллекта (ИИ) предоставляют огромные возможности для внедрения телемедицины в области здравоохранения.

В статье представлены экспериментальные результаты классификации медицинских данных для раннего выявления и предотвращения возможности возникновения СД2Т в системе принятия решений с использованием методов машинного обучения и интеллектуального анализа данных. Рассмотрена архитектура “умной” (smart) автоматизированной системы принятия медицинских решений для профилактики болезни СД2Т, которая разработана для выявления лиц, входящих в группу риска возможного развития данной болезни (prediabetes).

Проведен сравнительный анализ критериев эффективности работы рассмотренных методов ИИ для профилактики болезни СД2Т и оцениваются производительности следующих методов машинного обучения: метод опорных векторов (Support Vector Machine (SVM)), нейронные сети (Neural Networks (NNs)), логистическая регрессия (Logistic regression (LR)), наивный байесовский метод (Naпve Bayes (NBs)), случайный лес (Random Forest (RF)), деревья принятия решений (Decision Trees (DTs)) и метод к-ближайших соседей (k-Nearest Neighbors (K-NNs)). Рассматриваются различные показатели эффективности методов: Accuracy, F-score, Precision, Recall и ROC Area. Важным аспектом исследования является использование различных критериев эффективности для оценки методов машинного обучения с целью раннего выявления и предотвращения СД2Т. Исследование показало возможность получения высокой точности при анализе медицинских данных методами SVM, LR, NNs и RF.

Ключевые слова: интеллектуальный анализ данных, сахарный диабет 2-го типа, многоклассовая классификация, “умная” система принятия медицинских решений, машинное обучение с учителем, критерии эффективности.

Введение. Сахарный диабет классифицируется по двум основным клиническим категориям: диабет 1-го типа, или *“юношеский диабет”,* который является наследственным заболеванием (в основном болеют дети), и диабет 2-го

типа, или “диабет взрослых” [1,2]. В развитии болезни СД2Т большую роль играет сложное взаимодействие между факторами образа жизни и генетикой [2,3]. С возрастом риск развития СД2Т возрастает. Согласно среднестатистическим данным, этим заболеванием страдает каждый десятый житель планеты старше 40 лет. Диагностика болезни основана на выявлении повышенного уровня сахара крови у лиц с характерными признаками предрасположенности (ожирение, возраст, наличие диабета у близких родственников и т.д.) [4].

В настоящее время в области здравоохранения ведутся работы по разработке доступных, неинвазивных и надежных способов выявления риска развития СД2Т. Для прогностического выявления риска развития этой болезни широко используются разные электронные калькуляторы и шкалы риска (Diabetes Risk Score), из которых наиболее популярные DRS, AUSDRISK, FINDRISC. Недостатками способов раннего выявления и предотвращения болезни являются невозможность проведения долгосрочной профилактики СД2Т, а также необходимость использования дорогостоящих методов обследования, таких как HbA1c (гемоглобин), и сложность исполнения. В последние годы значительно возрос интерес к применению технологии машинного обучения для обработки большого объёма медицинских электронных данных [5]. Анализируя огромный массив накопленных медицинских данных и непрерывный поток данных мониторинга состояния здоровья в реальном времени, экспертные системы помогают поставить правильный диагноз СД2Т, разработать эффективную схему лечения. Использование методов машинного обучения обеспечивает высокую точность обнаружения, выявления и профилактики СД2Т, в результате чего, по мнению экспертов, в 2...3 раза сокращается не только стоимость лечения данной болезни, но и значительно улучшается качество лечения.

Постановка задачи. Целью работы является классификация медицинских данных с применением различных методов машинного обучения для оценки методов профилактики болезни СД2Т в “умной” системе принятия медицинских решений (УСПМР).

Структура системы. Рассматриваемая УСПМР имеет обобщенную структуру, которая приведена на рис. 1.

Рис.1. Структура УСПМР для профилактики СД2Т

В УСПМР процесс обработки медицинских данных организован на основе стандарта CRISP (Cross-Industry Standard Process for Data Mining), который представляет собой следующую последовательность этапов обработки больших данных (рис. 2): сбор, анализ и подготовка данных; первоначальный выбор параметров, нормализация данных; разработка модели; обучение и корректировка модели; тестирование модели; итеративное выполнение вышеуказанных шагов.

Рис.2. Схематическое изображение этапов моделирования в УСПМР

Набор данных. Исходные медицинские данные для профилактики болезни СД2Т:

1. Симптомы, семейная история и персональные данные, передаваемые с помощью мобильного приложения.
2. Тестовые данные пациента, поступающие от медицинских сенсоров (гаджетов).
3. Лабораторные данные, полученные от электронной медицинской базы данных.

В наборе данных каждая запись (instance) имеет следующую структуру: 12 входных атрибутов (x1-x12) и один выходной атрибут (y – результат диагностики)(см. табл.1).

Таблица 1

Атрибуты записи

Переменные атрибуты (Features)	Условие (Condition)	
x1 = HAlc (Hemoglobin (Alc))	x1 < 5.7 5.7 ≤ x1≤ 6.4 x1 > 6.4	класс 1 класс 2 класс 3
x2 = Cholesterol	< 200 mg 200≤ x2 ≤ 239 > 240 mg	нормальный пограничный высокий
x3 = Triglyceride	x3 < 150 mg 150 ≤ x3≤ 199 200 ≤ x3 ≤ 499 x3 > 500 mg	нормальный пограничный высокий очень высокий
x4 = BMI(weigth kg/height)	< 22 22 ≤ x4 ≤ 25 25 ≤ x4 ≤ 0 30 ≤ x4 ≤ 35 Fat > 35	худой нормальный избыточный толстый очень толстый
x5 = Blood Glucose (2hrs. p. p.)	x5 < 140 140 ≤ x5 ≤ 200 x5 > 200	класс 1 класс 2 класс 3
x6 = FBS(Fasting Blood Sugar)	x6 < 100 100 ≤ x6 ≤ 125 x6 >126	класс 1 класс 2 класс 3
x7 = age	x7 ≥ 40	
x8 = BP(blood pressure (mmHg))	x8 < 90/60 90/60 ≤ x8 ≤ 120/80 x8 > 120/80	низкий нормальный высокий
x9 = Family history of diabetes	x9 = 0/1 фактор наследственности	
x10 = Waist circumference	больше нормы, если x10 > 90 см окружность талии у мужчин и x10 > 85 см окружность талии у женщин	
x11 = Physical activity level	x11 = 0,1,2	высокий, средний, низкий,
x12 = Gender	x12 = 0/1	мужчина/женщина
Y = class(0,1,2)	класс 1 класс 2 класс 3	СД2Т отрицательный зона риска (prediabetes) СД2Т положительный

Классификация данных и сравнительный анализ производительности методов машинного обучения. Процесс обучения данных может производиться различными способами:

1) обучение с учителем (Supervised learning);

2) обучение без учителя (Unsupervised learning);

3) обучение с частичным привлечением учителя (Semi-supervised learning);

4) обучение с подкреплением (Reinforcement learning).

Для профилактики болезни СД2Т применяется метод обучения с учителем. При выборе метода обучения в УСПМР учитывается тот факт, что возможные ответы известны. Проблема раннего выявления риска заболевания СД2Т по результатам анализов медицинских данных представляет собой задачу многоклассовой классификации (рис. 3).

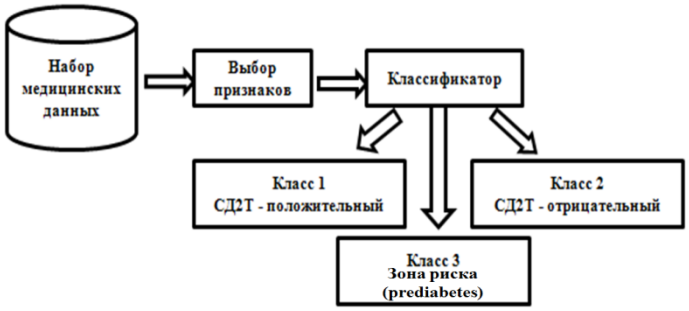

Рис.3. Результаты диагностики СД2Т

При обучении с учителем набор примеров X(samples) и правильных ответов решений Y формируется в некий алгоритм, задачей которого является найти некоторую функцию *f (X)*, преобразующую множество *X* в множество *Y:*

$$X \rightarrow f(X) \rightarrow Y. \quad (1)$$

Обучение с учителем является задачей обучения системы на тренировочном наборе данных. Путем приведения результатов обучения к тренировочному набору данных выявляются наиболее оптимальные параметры модели для прогнозирования возможных ответов на тестовых наборах данных, а с помощью набора валидации оценивается эффективность модели [6, 7]. На рис.4 изображена общая схема работы модели в системе.

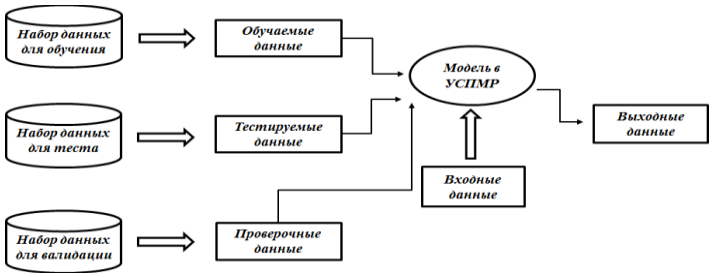

Рис.4. Общая схема работы модели в УСПМР

Для выбора методов модели учитываются следующие важные критерии:

- ***интерпретируемость*** - возможность объяснения, почему модель принимает именно это решение;
- ***простота*** - модель должна быть легко объяснимой и понятной;
- ***быстрота*** - производительность модели во время обучения и тестирования;
- ***масштабируемость*** - способность применения модели на больших наборах данных;

- ***точность модели*** - как правило, рекомендуется тестировать различные алгоритмы с различными параметрами и с помощью перекрестной проверки (cross-validation) сделать выбор модели, показывающей наилучшую производительность.

Для обучения и тестирования медицинских данных использовались следующие методы машинного обучения с учителем в среде Weka [8]: методы опорных векторов, наивный байесовский метод, метод к-ближайших соседей, логистическая регрессия, деревья решений, случайный лес, нейронные сети.

Для оценки качества алгоритмов классификации выбраны следующие численные метрики и показатели эффективности:

- ***True Positives (TP)*** - истинно положительные, количество пациентов в группе риска, диагноз положительный;
- ***False Positives (FP)*** - ложно положительные, количество здоровых пациентов, диагноз положительный;
- ***False Negatives (FN)*** - ложно отрицательные, количество пациентов в группе риска, диагноз отрицательный;
- ***True Negatives (TN)*** - истинно отрицательные, количество здоровых пациентов, диагноз отрицательный;
- ***Accuracy*** – правильность, доля правильных ответов:
  Accuracy = (TP + TN)/ (TP + TN + FP + FN);
- ***Precision*** – точность, оценка количества полученных от классификатора положительных ответов, являющихся правильными:
  *Precision = TP/ (TP + FP) *100;*
- ***Recall (R) (Sensitivity или True Positive Rate (TPR))*** – полнота, доля положительных ответов из ожидаемых:
  *Recall = TP/ (TP + FN) *100;*
- ***False Positive Rate (FPR)*** - показывает, какую долю из объектов отрицательного (negative) класса алгоритм предсказал неверно:
  FPR = FP/ (FP + TN);
- ***ROC Area*** - соотношение *TPR к FPR:*
  ROC/AUC = TPR / FPR;
- ***Sensitivity*** - вероятность того, что тест будет положительным при наличии риска заболевания:
  Sensitivity = TP / (TP + FN);
- ***Specificity*** – вероятность того, что тест будет отрицательным при отсутствии риска заболевания:
  Specificity = TN / (TN + FP);

Ошибки классификации - это False Negative (FN) и False Positive (FP) [9].

Точность классификации моделей для набора медицинских данных приведена в табл. 2.

Таблица 2

Результаты точности классификации моделей

Методы	Правильно классифицированные экземпляры %	Неверно классифицированные экземпляры %	TPR (Recall)	FPR	F-Measure	ROC Area	Precision
SVM	58.4699	41.5301	0.744 0.438 0.449	0.536 0.070 0.108	0.634 0.535 0.515	0.602 0.539 0.755	0.552 0.689 0.603
NBs	52.7322	47.2678	0.448 0.479 0.714	0.273 0.174 0.272	0.510 0.487 0.581	0.448 0.479 0.714	0.592 0.495 0.490
LR	59.2896	40.7104	0.680 0.521 0.510	0.433 0.115 0.127	0.627 0.565 0.549	0.664 0.833 0.790	0.582 0.617 0.595
K-NNs	45.9016	54.0984	0.488 0.375 0.490	0.500 0.204 0.172	0.476 0.385 0.500	0.513 0.615 0.651	0.464 0.396 0.511
DTs(J48)	45.3552	54.6448	0.506 0.427 0.388	0.505 0.215 0.164	0.487 0.421 0.422	0.471 0.633 0.587	0.470 0.427 0.463
RF (Макс. глубина 5)	54.6448	45.3552	0.703 0.323 0.490	0.524 0.564 0.600	0.600 0.411 0.539	0.703 0.323 0.490	0.605 0.786 0.776
NNs (3 скрытые слои)	57.6503	42.3497	0.657 0.500 0.510	0.448 0.119 0.134	0.608 0.545 0.543	0.611 0.776 0.707	0.565 0.600 0.581

От выбора атрибута зависит точность прогнозируемой модели. Наилучшую точность показали методы SVM, LR, NNs с тремя скрытыми слоями и метод RF с максимальной глубиной, равной 5.

Заключение. Исследования показали, что создание и интерпретация работы нейронной сети с использованием алгоритмов глубокого обучения происходят гораздо труднее, чем использование классификаторов RF, LR и SVM. Метод SVM отлично подходит для относительно небольших наборов данных, но требует достаточной настройки параметров. Вычислительная стоимость метода SVM растет линейно с количеством классов. Сложность метода RF возрастает по мере увеличения количества деревьев и примеров (samples) обучения. По сравнению с SVM, с помощью метода RF намного проще тренировать и легче получить надежную модель, кроме того, метод легче поддается распределенным вычислениям.

Эксперименты проводились на небольшом наборе медицинских данных. Для получения более точных результатов прогнозирования предполагается использование больших наборов медицинских данных.

Литература

1. American Diabetes Association. Classification and diagnosis of diabetes. - Diabetes Care 2015. - 38 Suppl: S8-S16.
2. http://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=80AEE25CE8B2275A8E6709EB95B043E9?sequence=1 .
3. Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins / **B. Newman, J.V. Selby, M.C. King, et al** // Diabetologia. – PubMed, 1987.- P.763–768.
4. Heritability estimates for beta cell function and features of the insulin resistance syndrome in UK families with an increased susceptibility to type 2 diabetes / **G.W. Mills, P.J. Avery, M.I. McCarthy, et al** // Diabetologia. – PubMed, 2004.- P.732–738.
5. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors / **N. Razavian, S. Blecker, et al** // Big Data.- 2015.- Vol.3, No.4.
6. https://medaboutme.ru/zdorove/spravochnik/bolezni/sakharnyy_diabet_2tipa /
7. **Witten I.H., Eibe F., Hall M.A.** Data Mining: Practical Machine Learning Tools and Techniques. - Morgan Kaufmann, San Francisco, 2011.- 558 p.
8. **Cleophas, Ton J., Zwinderman, Aeilko H.** Machine Learning in Medicine. - A Complete Overview, 2015.- 224 p.
9. Prediction models for risk of developing type 2 diabetes / **A. Ali, M.P. Linda, E. Corpeleijn, et al** // BMJ. - 2012.

Поступила в редакцию 02.04.2018.
Принята к опубликованию 05.06.2018.

ՄԵՔԵՆԱՅԱԿԱՆ ՈՒՍՈՒՑՄԱՆ ՄԵԹՈԴՆԵՐԻ ՀԵՆՔՈՎ ՏՎՅԱԼՆԵՐԻ ԴԱՍԱԿԱՐԳՈՒՄԸ ՈՐՈՇՈՒՄՆԵՐԻ ԿԱՅԱՑՄԱՆ ՀԱՄԱԿԱՐԳՈՒՄ 2-ՐԴ ՏԻՊԻ ՇԱՔԱՐԱՅԻՆ ԴԻԱԲԵՏԻ ԿԱՆԽԱՐԳԵԼՄԱՆ ՀԱՄԱՐ

Լ.Կ. Անդրեասյան

Առողջապահության համաշխարհային կազմակերպության վիճակագրության համաձայն՝ մեծահասակների ավելի քան մեկ երրորդը 2-րդ տիպի շաքարային դիաբետի (Տ2ՇԴ) զարգացման ախտանիշներ ունի: Տեղեկատվության և հեռահաղորդակցության տեխնոլոգիաների վերջին գիտական նվաճումները, ինչպես նաև IoT (Things of Intertnet) տեխնոլոգիաների լայնածավալ օգտագործումը հսկայական հնարավորություններ են տալիս առողջապահության ոլորտում հեռաբժշկության ներդրման համար:

Ներկայացված են մեքենայական ուսուցման և տվյալների իմացաբանական վերլուծության մեթոդների կիրառմամբ Տ2Շ վաղ հայտնաբերման և առաջացման հնարավորությունների կանխարգելման համար բժշկական տվյալների դասակարգման փորձարարական արդյունքները: Դիտարկվում է նաև Տ2Շ հիվանդության կանխարգելման համար խելացի բժշկական որոշումների կայացման համակարգի ճարտարապետությունը, որը նախատեսված է հիվանդության հնարավոր զարգացման ռիսկային գոտում (նախադիաբետ) գտնվող հիվանդների հայտնաբերման համար:

Տրվում է այդ հիվանդության կանխարգելման համար առաջարկվող ԱԲ ալգորիթմների աշխատանքի արդյունավետության չափանիշների համեմատական վերլուծությունը, և գնահատվում հետևյալ մեքենայական ուսուցման մեթոդների արագագործությունը. Support Vector Machine (SVM), Neural Networks (NNs), Logistic Regression (LR), Naive Bayes (NBs), Random Forest (RF), Decision Tree (DTs) և K-nearest Neighbors (KNNs): Դիտարկվում են ալգորիթմների արդյունավետության տարբեր չափանիշներ. Accuracy, F-score, Precision, Recall և ROC Area:

Գիտական հետազոտության կարևոր հայեցակետ է S2Շ հիվանդության կանխարգելման և վաղ հայտնաբերման համար կիրառվող մեքենայական ուսուցման մեթոդների գնահատման տարբեր արդյունավետության չափանիշների օգտագործումը: Հետազոտության արդյունքները ցույց են տվել բարձր ճշգրտություն ստանալու հնարավորությունը SVM, LR, NNs և RF մեթոդների կիրառման դեպքում:

Առանցքային բառեր. տվյալների ինտելեկտուալ վերլուծում, 2-րդ տիպի շաքարային դիաբետ, բազմադասային դասակարգում, բժշկական որոշումների կայացման խելացի համակարգ, վերահսկելի մեքենայական ուսուցում, արդյունավետության չափանիշներ:

CLASSIFICATION OF DATA IN THE DECISION-MAKING SYSTEM FOR THE PREVENTION OF TYPE 2 DIABETES BASED ON THE METHODS OF MACHINE LEARNING

L.K. Andreasyan

According to the statistics of the World Health Organization, more than one-third of the adult population have symptoms of developing type 2 intellectual data analysis of diabetes mellitus (T2DM). Scientific advances in information and telecommunication technologies, the development of IoT technologies (Intertnet of Things) and the widespread use of artificial intelligence (AI) methods provide tremendous opportunities for the introduction of telemedicine in the field of health.

The article presents the experimental results of the medical data classification for the early detection and prevention of the occurrence possibility of T2DM in the decision making system, using the methods of machine learning and intellectual data analysis. The article also examines the architecture of the smart medical decision making system for the prevention of T2DM disease, which is designed to identify individuals at a risk of a possible development of the disease (prediabetes).

The article compares the efficiency criteria of the considered AI methods for the prevention of T2DM disease and evaluates the performance of the following methods of machine learning: Support Vector Machine (SVM), Neural Networks (NNs), Logistic Regression (LR), Naive Bayes (NBs), Random Forest (RF), Decision Trees (DTs) and k-Nearest Neighbors (K-NNs). Various performance indicators of methods are considered: Accuracy, F-score, Precision, Recall and ROC Area. An important aspect of the study is the use of various performance criteria to evaluate machine learning methods for early detection and prevention of T2DM. The study proves the feasibility of obtaining high accuracy in the analysis of medical data by the methods SVM, LR, NNs and RF.

Keywords: intellectual data analysis, type 2 diabetes mellitus, multiclass classification, smart medical decision making system, supervised machine learning, performance criterias.