КАК „ТИТАНИК“ ПРОДЪЛЖАВА ДА ПРОМЕНЯ БЪДЕЩЕТО - МАШИННО ОБУЧЕНИЕ В РЕСПИРАТОРНАТА ПРАКТИКА

Евгени Меков[1], Росен Петков[2]

[1]Катедра по професионални болести, Медицински факултет, Медицински университет - София
[2]Катедра по белодробни болести, Медицински факултет, Медицински университет - София

HOW TITANIC CONTINUES TO CHANGE THE FUTURE: MACHINE LEARNING IN RESPIRATORY PRACTICE

Evgeni Mekov[1], Rosen Petkov[2]

[1]Department of Occupational Diseases, Faculty of Medicine, Medical University of Sofia
[2]Department of Respiratory Diseases, Faculty of Medicine, Medical University of Sofia

РЕЗЮМЕ

Приложението на изкуствен интелект (ИИ) във всички области и в частност – в медицината, придобива все по-голяма популярност. Машинното обучение (МО) е част от ИИ, при което компютрите използват статистически методи за самообучение, без да са изрично програмирани за това. Предизвикателството „Titanic - Machine Learning from Disaster“ е вероятно най-популярното за изследователите, навлизащи в сферата на изкуствения интелект и машинното обучение.

В почти всички случаи МО преминава през няколко последователни етапа: дефиниция на проблема, получаване на данни; подготовка на данните; изследователски анализ; създаване на характеристики; създаване на модел; оценка на модела. Тази статия описва накратко всеки от тях. С тяхна помощ е създаден модел за предизвикателството „Titanic - Machine Learning from Disaster“, който е по-добър от повечето предложени модели в състезанието (постига по-добър резултат от 99.8% от моделите).

Машинното обучение може да подобри точността на различните модели при едни и същи данни (равен старт). По този начин правилната обработка на данните може да доведе до по-висока точност, което в медицината се изразява в подобрение на клинично значими крайни точки.

Ключови думи: *машинно обучение, изкуствен интелект, алгоритъм, точност, „Титаник“, медицина*

ABSTRACT

The application of artificial intelligence (AI) in all fields and in medicine in particular is gaining more and more popularity. Machine learning (ML) is part of AI, in which computers use statistical methods for self-learning without being explicitly programmed to do so. The challenge Titanic: Machine Learning from Disaster is probably the most popular among researchers entering the field of artificial intelligence and machine learning.

In almost all cases, ML goes through several successive stages: definition of the problem, data gathering, data preparation, exploratory analysis, feature engineering; model creation, and model evaluation. The following manuscript briefly describes each of them. With their help, a model for the challenge „Titanic: Machine Learning from Disaster“ was created. This model outperformed most of the models featured in the competition. It achieved a 99.8% better result than the other models presented.

Machine learning could improve the accuracy of different models with the same data (equal start). In this way, proper data processing could lead to higher accuracy, which in medicine translates into improved clinically significant endpoints.

Keywords: *machine learning, artificial intelligence, algorithm, accuracy, titanic, medicine*

ВЪВЕДЕНИЕ

Приложението на изкуствен интелект (ИИ) във всички области и в частност – в медицината, придобива все по-голяма популярност. Машинното обучение (МО) е част от ИИ, при което компютрите използват статистически методи за самообучение, без да са изрично програмирани за това (1).

Предизвикателството „Titanic - Machine Learning from Disaster“ е вероятно най-популярното за изследователите, навлизащи в сферата на изкуствения интелект и машинното обучение (2). То изисква създаване на модел, използвайки машинно обучение, който прогнозира кои пътници са оцелели от корабокрушението на „Титаник”.

В почти всички случаи МО преминава през няколко последователни етапа. Тази статия описва накратко всеки от тях.

Историята на „Титаник“

Потъването на „Титаник“ е едно от най-ужасяващите корабокрушения в историята. На 15 април 1912 г. смятаният за „непотопяем“ RMS (Royal Mail Ship, Кралски пощенски кораб)

„Титаник“ потъва след сблъсък с айсберг. За съжаление няма достатъчно спасителни лодки за всички на борда, което води до смъртта на 1502-ма от 2224-ма пътници и екипаж.

Въпреки че оцеляването е свързано до известна степен и с късмет, някои групи хора имат по-висока вероятност да оцелеят от други. В предизвикателството „Titanic - Machine Learning from Disaster“ участниците трябва да създадат модел, който да отговори на въпроса „кой тип хора е по-вероятно да оцелее?“, като използват данните за пътниците (например име, възраст, пол, социално-икономически клас и т.н.).

Алгоритъм за машинно обучение

Основните етапи на МО са:

- Дефиниция на проблема;
- Получаване на данни;
- Подготовка на данните;
- Изследователски анализ;
- Създаване на характеристики;
- Създаване на модел;
- Оценка на модела.

Налични данни

В това състезание са налични два сходни набора от данни, съдържащи пътническа информация (табл. 1). Тренировъчният набор от данни съдържа данни за 891 пътници, заедно с реалната информация за преживяемостта на всеки от тях (т. нар. фундаментална истина). Тестовият набор от данни съдържа подобна информация за други 418 пътници, но без да разкрива фундаменталната истина за преживяемостта им. Именно последната е параметърът, който трябва да се прогнозира, използвайки установени от участниците зависимости в тренировъчния набор.

Табл. 1. Променливи в „Titanic - Machine Learning from Disaster“

Променлива	Описание
Survived	Оцелял (1) или умрял (0)
Pclass	Пътническа класа
Name	Име
Sex	Пол
Age	Възраст
SibSp	Брой на братя/сестри и съпрузи на кораба
Parch	Брой на родители/деца на кораба
Ticket	Номер на билета
Fare	Цена на билета
Cabin	Номер на каютата
Embarked	Пристанище на качване

Подготовка (почистване) на данни (data clean)

Подготовката на данните е в основата на машинното обучение. Тя включва почистване на данните и създаване на характеристики. Познаването на областта (т. нар. domain knowledge) също е от голяма важност за постигане на добри резултати. Подготовката на данните не може да бъде напълно автоматизирана или поне не в началото. Тя отнема огромната част от времето (60-80%) за създаване на алгоритъм, но е задължителна част за постигане на възможно най-добра точност.

Подготовката на данните представлява процес на откриване и коригиране (или премахване) на грешни или неточни записи в набор от данни и се отнася до установяване/идентифициране на непълни, неправилни, неточни или неподходящи стойности и след това тяхното заместване, модифициране, или изтриване.

Изчистването на данните задава тяхната правилна форма и качество. То включва много функции като например филтриране на подгрупи, премахване на повторения, трансформация (нормализация, стандартизация), прегрупиране, замяна на данни, запълване на липсващите данни. Откритите или отстранени несъответствия често са причинени от грешки при въвеждането на данни. След почистване наборът от данни трябва да е в съответствие с други сходни данни.

Качеството на данните може да се оцени с помощта на няколко критерия:

- валидност – степента на съответствие на измерванията със зададени правила или ограничения;
- точност – степента на съответствие на едно измерване спрямо зададен стандарт или истинската стойност;
- пълнота – степента, до която са известни всички необходими измервания;
- съгласуваност (последователност) – степента на еквивалентност на измерванията в проучването спрямо други проучвания;
- еднородност – употребата на еднакви мерни единици между различни проучвания.

Изследователски анализ и създаване на характеристики (Feature engineering)

Създаването на характеристики е процесът на използване на познанията в специфичната област за получаване на допълнителни характеристики от непреработени данни. Тези характеристики могат да се използват за подобряване на ефективността на алгоритмите за машинно обучение. Създаването на характеристики (feature engineering) се състои от анализ на параметрите и избор на правилните компоненти за анализ, валидиране на параметрите и подобрения на избраните параметри. Процесът по създаване на характеристики включва две стъпки:

- трансформация на променливи – замяна на променлива от функция (например коренуване на променливата, логаритмуване на променливата и т.н.). Това е процес на промяна на разпределението или зависимостта на променливата с други променливи;
- създаване на нови променливи/характеристики – за допълнително подобряване на зависимостта между променливите.

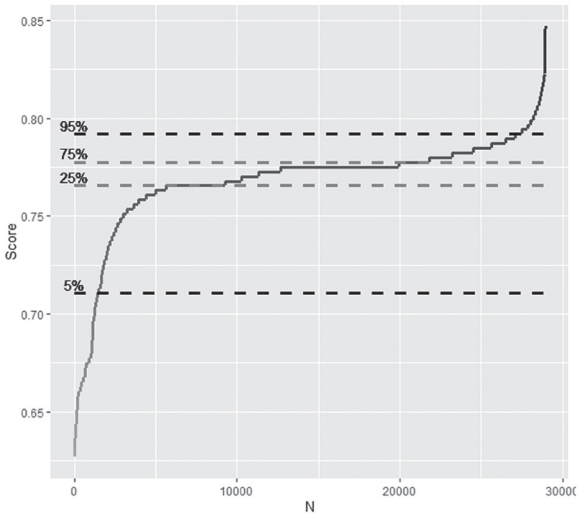
Създаване на модел

За да се оцени качеството на създадените модели в „Titanic - Machine Learning from Disaster“, трябва да се знае т. нар. нулев модел – модел без предиктори. Модел, който предсказва, че всички умират, има точност 62.7% (табл. 2). С оглед спецификата на спасителната операция полът е значим предиктор за преживяемостта. Базов модел, който предсказва, че всички мъже умират, а всички жени оцеляват, значимо подобрява точността - 76.6% (табл. 2).

С помощта на модел, който включва почистване на данните (заместване на липсващите стойност със средната стойност или медиана, премахване на ненужните данни), изследователски анализ, създаване на характеристики и настройка на хиперпараметрите се постига точност 82.057% (3). Този модел заема място в първите 0.2% (60/28930), ако се премахнат моделите с точност <0.626 (нулев модел) и >0.84689 (моделът с най-висока известна потвърдена точност). Трябва да се отбележи, че 50% от моделите попадат в интервал с разлика от точността 1.2% (76.55% - 77.75%), т.е. минимална промяна в точността на модела може да определи подобрението му от четвъртия в първия квартил (фиг. 1).

Табл. 2. Резултат от различните модели

Модел	**Точност**
Всички умират	62.7%
Мъжете умират, жените оцеляват	76.6%



Фиг. 1. Разпределение на резултатите

Настройка на хиперпараметрите

Хиперпараметърът представлява параметър, чиято стойност се използва за контрол на процеса на обучение. В машинното обучение оптимизацията или настройката на хиперпараметрите представлява избора на набор от оптимални хиперпараметри за алгоритъма. Настройването на хиперпараметрите позволява допълнително увеличаване на точността на модела (4). За разлика от тях стойностите на другите параметри (най-често тежестта на информиращите възли) се определят в самия процес на обучение.

Ансамблови методи

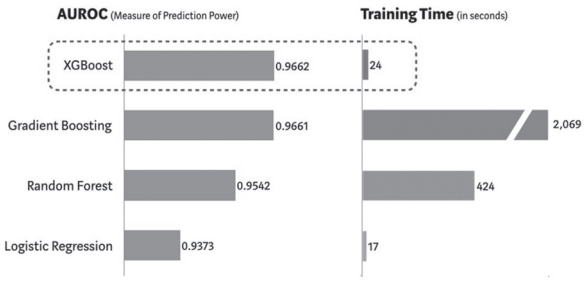
Ансамбловите методи (ensemble methods) комбинират статистика и МО, като използват съвкупност от обучаващи се алгоритми, за да постигнат по-добра предиктивна точност в сравнение с всеки от съставящите ги алгоритми поотделно (5,6).

При неструктурирани данни (изображения, текст и т.н.) изкуствените невронни мрежи са златен стандарт спрямо всички други алгоритми. Въпреки това, когато става въпрос за малки до средни структурирани/таблични данни, алгоритмите, базирани на дървото на решенията (decision tree), към момента постигат най-добри резултати. XGBoost е алгоритъм за машинно обучение, базиран на дърво на решенията, който използва градиентно усилване (7). Еволюцията на XGBoost алгоритъма е представена на табл. 3.

Табл. 3. Развитие на метода XGBoost

Метод	**Основни характеристики**
Дърво на решенията (Decision tree)	Графично представяне на възможните сценарии при решението, основано на определени условия
Bagging	Ансамблов метаалгоритъм, комбиниращ прогнозите от множество дървета на решенията чрез мажоритарно гласуване
Random forest	Производен на Bagging алгоритъм, при който част от характеристиките са случайно избрани за формиране на „гора“ – набор от дървета на решенията
Усилване (Boosting)	Моделите се развиват последователно, като намаляват грешките от предходните модели и усилват въздействието на моделите с добри резултати
Градиентно усилване (Gradient boosting)	Използва алгоритъм с постепенно усилване, който минимизира грешките в последващите модели
XGBoost	Оптимизиран gradient boosting алгоритъм чрез паралелна обработка, окастряне на дърветата на решенията, обработка на липсващите данни и регуларизация, с цел избягване на преобучение/грешка

XGBoost значително превъзхожда други алгоритми при еднакви данни по отношение на точност и необходимо време за създаване на модела (фиг. 2).



Фиг. 2. Точност и тренировъчно време на XGBoost срещу други ML алгоритми. Случайна извадка от 1 милион наблюдения с 20 характеристики (2 информативни и 2 излишни) (по (7).

Оценка на модела

Оценката на модела има за цел да оцени точността на даден модел върху бъдещи (непознати/извън извадката) данни. Изборът на параметрите за оценка зависи от конкретната задача (напр. класификация, регресия, класиране, клъстеризация, моделиране и т.н.). Основните показатели за оценка на класификационен модел са точност (accuracy), прецизност (precision) и пълнота (recall).

ДИСКУСИЯ

От създаването на регресионния модел през 1805 г. процесът на обработка на данни непрекъснато се усъвършенства. Към момента boosting алгоритмите са златен стандарт, печелейки огромната част от състезанията и предизвикателствата за МО. Машинното обучение е обект на активна изследователска дейност, като дори вече са налице няколко алтернативи на XGBoost. LightGBM (Microsoft) показва голям потенциал (8). CatBoost (Yandex) вече се налага като водещ модел при преобладаващи категорийни данни, като в някои области изпреварва конкурентите си (фиг. 3) при необходимо време за създаване на модела под 10% от това на останалите (9).

	CatBoost	LightGBM		XGBoost		H2O	
Adult	0.269741	0.276018	+2.33 %	0.275423	+2.11%	0.275104	+1.99%
Amazon	0.137720	0.163600	+18.79 %	0.163271	+18.55%	0.162641	+18.09%
Appet	0.071511	0.071795	+0.40 %	0.071760	+0.35%	0.072457	+1.32%
Click	0.390902	0.396328	+1.39 %	0.396242	+1.37%	0.397595	+1.71%
Internet	0.208748	0.223154	+6.90 %	0.225323	+7.94%	0.222091	+6.39%
Kdd98	0.194668	0.195759	+0.56 %	0.195677	+0.52%	0.195395	+0.37%
Kddchurn	0.231289	0.232049	+0.33 %	0.233123	+0.79%	0.232752	+0.63%
Kick	0.284793	0.295660	+3.82 %	0.294647	+3.46%	0.294814	+3.52%

Logloss

Фиг. 3. Сравнение на точността (Logloss) на CatBoost, LightGBM, XGBoost и H2O при различни данни (по (9)).

Осъществяване на информирана клинична преценка въз основа на съществуващите дан-

ни е в основата на медицината, базирана на доказателствата (10). В последните години МО навлиза с бързи темпове и в медицината. Някои от най-значимите и актуални области включват:

- прогнозиране на разпространението на COVID-19 (11);
- установяване на полуживота на ваксините срещу CWOVID-19 (12);
- установяване на позицията на ендотрахеална тръба или катетър (13);
- установяване на локализацията и вида на рентгенологични изменения (14).

ИИ и МО могат да доближат медицината до персонализиран подход, като предоставят прогноза/риск за конкретния пациент, основани на неговите характеристики (индивидуален подход). Това е стъпка напред спрямо често използваното към момента причисляване на пациента към определена група, за която има данни за прогнозата/риска (групов подход). Важно е да се отбележи, че при наличието на едни и същи данни (равен старт) различните модели имат различна точност. По този начин обработката на едни и същи данни с най-уместния модел може да доведе до по-висока точност и в крайна сметка до подобрение на клинично значимите крайни точки.

ЗАКЛЮЧЕНИЕ

Приложението на изкуствен интелект (ИИ) във всички области и в частност – в медицината, придобива все по-голяма популярност. Машинното обучение е част от ИИ, при което компютрите използват статистически методи за самообучение, без да са изрично програмирани за това. В почти всички случаи МО преминава през няколко последователни етапа: дефиниция на проблема; получаване на данни; подготовка на данните; изследователски анализ; създаване на характеристики; създаване на модел и оценка на модела.

Машинното обучение може да подобри точността на различните модели при едни и същи данни (равен старт). По този начин правилната обработка на данните може да доведе до по-висока точност, което в медицината се изразява в подобрение на клинично значими крайни точки.

ЛИТЕРАТУРА

1. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Springer; 2013. Available from: https://faculty.marshall.usc.edu/gareth-james/ISL/
2. https://www.kaggle.com/c/titanic
3. https://www.kaggle.com/medaxone/titanic-challenge-0-82057-for-beginners-with-tune
4. Claesen M, De Moor B. Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127, 2015.
5. Rokach L. Ensemble-based classifiers. Artificial Intelligence Review 2010;33(1–2):1–39.
6. Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 2006;6(3):21–45.
7. https://xgboost.readthedocs.io/en/latest/
8. https://lightgbm.readthedocs.io/
9. https://catboost.ai/
10. Buch V. Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract. 2018;68(668):143–144.
11. https://www.kaggle.com/c/covid19-global-forecasting-week-5
12. https://www.kaggle.com/c/stanford-covid-vaccine
13. https://www.kaggle.com/c/ranzcr-clip-catheter-line-classification
14. https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection

Адрес за кореспонденция:
Доц. д-р Евгени Меков, д.м.
УМБАЛ „Св. Иван Рилски“ ЕАД
бул. „Акад. Иван Гешов“ 15
София, 1000
e-mail: Evgeni.mekov@gmail.com