Подготовка на данни за машинно обучение и изкуствен интелект и складове от данни

Data Preparation for Machine Learning and Artificial Intelligence, and Data Warehouses

Генка Митева[1]

Абстракт

Складовете от данни са вече добре установена система за съхранение на данни, които в последствие се използват за анализ, като събират данни от много различни източници след тяхната трансформация. Те предлагат възможност за автоматизацията на процеса по подготовка на данни, който включва няколко основни стъпки.
При подготовка на данни за машинно обучение също се преминава през подобни етапи. Моделите за машинно обучение обаче изискват малко по-специфични набори от данни. Сходствата в двата процеса обаче правят възможно създаването на система, автоматизираща подготовката на данни чрез разширяване на архитектурата на склада от данни.

Abstract

Data warehouses are an already well-established system for the storage of data, which is later used for analysis, which contains data from many different sources after transforming them. They offer an opportunity to automate the data preparation process, which includes a few main steps.
The data preparation process for machine learning also includes similar stages. The machine learning models need datasets with some specific characteristics. However, the similarities in these two processes facilitate the creation of a system which automates the data preparation process by expanding the data warehouse architecture.
Ключови думи: Машинно обучение, складове от данни, подготовка на данни

Въведение

В днешно време организациите все повече разчитат на получаването на нова информация от наличните им данни. Складовете от данни спомагат в интегрирането и управлението на голямото количество информация от различни източници. Ключова характеристика на съвременните складове от данни е способността им да автоматизират основите задачи по подготовката на данни, като почистване, трансформация и т.н., които са основни при подготовката на данни за анализ.

Архитектурата на склада от данни полага добра основа, върху която да се надградят няколко допълнителни стъпки. Идеята е да бъдат включени и стъпките от подготовката за данни за машинно обучение, които в момента липсват в традиционната архитектура на склада от данни чрез добавянето на различни процеси като подбор на характеристики, разделяне на данни и т.н. По този начин складът от данни може да послужи за цялостна система, която автоматизира трудния и скъп процес по подготовка на данни за машинно обучение. По този начин проектите, свързани с машинно обучение могат да бъдат оптимизирани и ускорени, намалявайки нуждата от ръчна обработка на

[1] Докторант, катедра ИТК, УНСС, София, e-mail: gmiteva@unwe.bg

данните. Възможно е и да се намалят грешките, причинени от нуждата от ръчна обработка, което да направи моделите по-точни и надеждни. Складовете от данни могат да играят решаваща роля в проекти, свързани с машинно обучение. При добре структуриран склад от данни, организацията може да си осигури работа с актуални, консистентни данни.

Събирането на данните и проблемите с качеството им са важни, но до момента голяма част от литературата на тема машинно обучение се фокусира основно върху обучаването на алгоритми, а не върху самите данни. Докладът цели да направи обобщение на стъпките от процеса по подготовка на данни, както и да изследва тяхното място в архитектурата на склада от данни, която да бъде допълнена с цел създаване на система за подготовка на данни за машинно обучение.

Подготовка на данни

Подготовката на данни е процесът по обработка на сурови данни, така че да са подходящи за по-нататъшна обработка и анализ. Основните стъпки включват събиране, почистване, етикетиране и т.н. Подготовката на данните може да отнеме до 80% от времето за работа върху проект за машинно обучение. Поради тази причина, автоматизирането на процеса може да е решаващо за оптимизирането на процеса. Използването на данните на организацията може спомогне за вземането на по-информирани решение и откриването на нови, неочаквани възможности. Това е важен, но тежък процес, който е решаващ за постигането на точен модел за машинно обучение. За да се минимизира времето, отделено за тази част от проекта могат да бъдат използвани различни инструменти, които да автоматизират подготовката на данни.

Събиране на данни

Целта на събирането на данни е да бъдат открити набори от данни, които да бъдат използвани за обучаване на моделите за машинно обучение. В някои източници, тук спада и обогатяването на данни, заедно с откриването на данни и генерирането на данни. Откриването на данни (data discovery) е нужно в ситуации, в които се споделя или търси нов набор от данни и тази стъпка става все по-важна, имайки предвид количеството данни, което имаме на разположение. Съществуват много системи, създадени за колаборация, които улесняват процеса. Други системи обаче нямат за цел споделянето на набори от данни. При тях метаданните се генерират след създаването на набора от данни, без помощта на създателя му. Основно предизвикателство в подобни системи е липсата на сигурност, че наборът от данни е подходящ за съответния проблем.

Техниките за събиране на данни за машинно обучение са решаващи при правенето на проучване или взимането на решение в организацията, тъй като това е една от стъпките, осигуряващи качеството на данни. На този етап може да осигурим надеждни данни, които да спомогнат в процеса по откриване на знание.

Има два основни метода за събиране на данни:

- Първично събиране на данни - данните са нови и оригинални, събрани директно от източника и не са били използвани преди това. Информацията, получена чрез този тип техники е точна и

събрана конкретно за целите на съответното проучване. Тя може да е събрана от проучвания, въпросници, наблюдения, сензори, експерименти, фокус групи

- Вторично събиране на данни - данните вече са били използвани, вторични данни. Данните могат да бъдат транзакционни данни от различни системи, данни от публични бази данни, активност в социални мрежи, събиране на информация от уеб сайтове чрез различни приложения.

Генерирането на данни може да се използва когато няма подходящи външни източници на данни, но е възможно да се генерира набор от данни по изкуствен начин. Тази техника опитва да се справи с проблеми като защитата на лични данни и сигурността на данните.

Почистване на данни

Обучаването на модели върху все по-големи количества данни е ключово предизвикателство при управлението на данни. Въпреки че има много рамки на обучение, които покриват доста от основните стъпки по процеса на машинно обучение, те рядко предлагат решения за това кои характеристики да се използват или как да се представят данните. Създаването на модел е повтарящ се процес, който често разчита на проба и грешка. Фактът, че данните невинаги са готови за обработка допълнително усложнява работата, като в данните често има липсващи, грешни или неконсистентни стойности, които може да се дължат на повредени сензори, софтуер, хардуер. Почистването на данните може да включва няколко основни задачи:

- Откриване на извънредни стойности
- Премахване на грешни стойност
- Премахване на дублирани стойности
- Справяне с липсващи стойности

При почистването на данни е невъзможно да бъде открито решение за автоматизация, което да сработва във всички случаи и при всички набори от данни. В различен контекст могат да бъдат открити различни проблеми.

Трансформация на данни

Трансформацията на данни се състои в конвертирането и оптимизирането на данни за различни цели, които може да включват анализ, изработка на доклади или съхранение. Целта на процеса е да направи данните по-достъпни, разбираеми и удобни за обработка, така че да спомогнат взимането на информирани решения.

Процесът по трансформация на данните може да включва следните стъпки:

- Кодиране на категорийни данни
- Стандартизация
- Агрегиране на данни
- Създаване на характеристики
- Заглаждане
- Предварителна обработка на текст

Трансформацията на данни е решаваща стъпка в анализът на данни и машинното обучение, която може значително да повлияе на представянето на модела. Трансформацията на данните ги прави по-консистентни, намалява шума, спомага за използването на пълния потенциал на наличните данни.

Етикетиране на данни

При машинното обучение, етикетирането на данни е процесът по идентифициране на сурови данни (изображения, текстови файлове, видео) и добавянето на етикети или маркери, които носят контекст. В момента повечето модели за машинно обучение са контролирани (supervised). За работа с този тип модели е нужен набор от данни, който е маркиран, така че моделът да се обучен върху него. Етикетирането на данни обикновено започва с процес, в който хора вземат решения за неетикетирани данни.

Няколко често срещани вида етикетиране на данни са:

- Компютърно зрение
- Алгоритми за обработка на естествен език
- Обработка на аудио

За успешното обучаване на модел са нужни големи количества данни за трениране. Откриването на тези данни обаче често е скъпо, сложно и времеемко. Голяма част от моделите в днешно време се нуждаят от човешка намеса за етикетирането на данни.

Намаляване на размерността

Намаляването на размерността е една от най-популярните техники за премахване на шум от данни, както и премахването на ненужни характеристики. Обикновено има два типа намаляване на размерност - подбор на характеристики и отделяне на характеристики.

Подбор на характеристики

Подборът на характеристики е ефективна стъпка в подготовката на данни, особено многомерни такива, за различни проекти за машинно обучение и извличане на закономерности от данни. Целта на този етап е да се построи по-прост и разбираем модел, като подобри представянето му и подготви чисти и разбираеми данни.

Подборът на характеристики подобрява представянето на модела и неговата изчислителна ефективност, като се премахват ненужните характеристики. Идеята на процеса е да избере малка група от релевантни характеристики от целия набор според критерий, по който се оценява кои характеристики са най-подходящи и водят до най-добро представяне на модела. В зависимост от това дали тренировъчният набор от данни е етикетиран или неетикетиран, моделите за подбор на характеристики могат да бъдат контролирани, неконтролирани и полуконтролирани.

Контролираните методи могат да бъдат разделени на филтър модели, обгръщащи модели и вградени модели. Филтър моделите разчитат на измерването на основните характеристики на обучителните данни като разстояния, консистентност, зависимост, информация и корелация.

Обгръщащите методи използват точността на прогнозиране на модела, за да оценят качеството на избраните характеристики.

Вградените методи се опитват да предложат решение на недостатъците на предходните два типа модели , като използват статистически модели, както филтър методите, за да избере няколко групи от характеристики. След това, избира групата с най-голяма точност на класификация, като по този начин постига точността на обгръщащите модели с изчислителната ефикасност на филтър моделите.

Разделяне на данните

Разделянето на данните е процес, която се смята за задължителен при машинното обучение, за да се елиминират или намалят систематичните грешки в моделите за машинно обучение. Възможно е данните да бъдат разделени на две части - за обучение и тестови. Най-често обаче идеята е пълният набор от данни да бъде разделен на данни за обучение, данни за тест и данни за валидация. Тази част от процеса е задължителна за обучаването на модела, настройване на параметрите му и оценяване на представянето му.

Най-често се използва крос-валидация - данните се разделят на две подгрупи, като един се използва за обучаване на модела, а другата - за оценяване на представянето му. Основната цел на този метод е да се постигне стабилна оценка на представянето на модела. Двата по-често използвани вида cross-validation са hold-out и k-fold. При hold-out крос-валидацията процесът е ефективен и лесен. Данните се делят на трите вече споменати подгрупи, като пропорцията на разделяне не е строго ограничена. При k-fold крос-валидацията, се използва комбинация от повече тестове, за да се оцени грешката на модела. Това може да се окаже полезно в ситуации, при които няма достатъчно данни за hold-out крос-валидация. Данните се разделят на k на брой равни части. Една от тях се използва за валидация (тестване), а останалите - за обучаване. Процесът се повтаря за всяка подгрупа.

Проблемът с правилното разделяне на данни може да се реши като статистически проблем за намирането на представителна извадка чрез следните методи:

- Проста случайна извадка
- Стратифицирана случайна извадка
- Систематична случайна извадка и др.

Обогатяване на данни

Терминът обогатяване на данни означава повтарящият се процес по оптимизация на алгоритми чрез представянето на непознати данни. Когато за обучаването на даден модел се използват голямо количество данни, представящи много различни възможни случаи, това подобрява представянето на модела. Този процес може да бъде използван в случаи, в които липсват достатъчно данни за целите на проекта или целим да подобрим представянето на модела като му представим допълнителна информация.

Чрез обогатяването на данните изкуствено се генерират нови данни, като се правят промени по оригиналните такива. Освен че подобрява представянето на модела и намалява нуждата от откриването на още данни, което е труден и скъп процес, обогатяването на данни спомага и за

преодоляване на преобучението - ситуация, в която моделът твърде много се базира върху обучителните данни и не обобщава добре при представянето на нови такива.

Някои основни техники за обогатяване на данни са компютърно зрение, обогатяване на аудио данни, обогатяване на текстови данни.

Същност и архитектура на склада от данни

Складовете от данни са може би най-често използваната технология за управление и извличане на информация от данни в организацията. Те съхраняват огромни количества данни, което ги прави подходящи и за използването им в машинното обучение. Складовете от данни имат няколко основни функции - извличане на данни от различни източници, почистване, подготовка, зареждане и поддръжката им, обикновено в релационна база данни. Една от често срещаните архитектури, която би била подходяща и за подготовка на данни за машинно обучение е трислойната:

- Източници - това обикновено е първият слой на склада от данни, където се извличат данни то различни източници. В повечето организации има голямо количество и разнообразие от източници, с различни модели и формат на данните. Това могат да бъдат различни файлове, системи за управление на бази данни, XML документи и т.н.
- Междинна част - междинната част на склада от данни може да съдържа различни инструменти, чиято цел е да филтрират и трансформират данните от различните източници, така че да са подходящи за съхранение. В този слой данните се съхраняват временно, като задачите, които се изпълняват тук са от изключителна важност за качеството и консистентността на данните. Огромното количество възможни формати на данните налагат те задължително да бъдат обработени, за да бъдат съхранени и подходящи за по-нататъшен анализ. В този слой се изпълняват задачи като почистване, трансформация, нормализация и интегриране на данните от различни източници, така че да бъдат съхранени в едно централизирано хранилище. На това ниво данните обикновено се трансформират в OLAP куб, в който се съхраняват различни разрези на данните.
- Аналитичен слой - този слой предоставя възможност зса анализ, изработка на доклади и визуализиране на данните. След като суровите данни са били обработени и трансформирани, тук вече те са на разположение за използване от различни системи за бизнес интелигентност. Данните могат да бъдат разпределени в модули данни според специфичната сфера от бизнеса, за която се отнасят. Тук могат да бъдат автоматизирани процесите по генериране на доклади и генерирането на различни данни, нужни на бизнеса.

Стъпки от подготовката на данни в архитектурата на склада от данни

Тази архитектура е подходяща и за извършване на част от стъпките, които трябват да бъдат изпълнени при подготовка на данни за машинно обучение. Някои от тях вече присъстват в архитектурата на традиционния склад от данни.

- **Събиране на данни** - в традиционния склад от данни, тази стъпка се изпълнява на първия слой при извличането на данни от различни източници.

- **Почистване на данни** - тази стъпка също присъства в архитектурата на склада от данни в повечето случаи. При трансформацията на данни за използването и съхранението им в склад от данни често се налага да се обработват различни липсващи или крайни стойности, както и да се премахват дублирани данни и характеристики.
- **Трансформация на данни** процесът по трансформация на данните също често се налага да бъде изпълнен в склада от данни, обикновено отново в междинния слой, като почистените данни трябва да бъдат преформатирани в подходяща структура. Този прозес може да включва агрегираните, сортиране, кодиране, нормализация. След това, трансформираните данни се зареждат в склада, което позволява ефективната им обработка и анализ.

Допълнителни стъпки в подготовката на данни

Някои от стъпките по подготовка на данни не се изпълняват в склада от данни, тъй като те са специфични за машинното обучение. Архитектурата на традиционни склад от данни може да бъде допълнена, за да се добавят и автоматизират останалите стъпки от процеса, които включват:

- Подбор на характеристики - тази част от процеса по подготовка на данни за машинно обучение е решаваща за доброто представяне на модела и би следвало да е включена в идеята за система, автоматизираща такива задачи. Възможно е на базата на алгоритми за подбор на характеристики да се съхраняват метаданни с най-подходящите характеристики за определени цели и за работа с конкретни алгоритми. Това би улеснило последващото използване на данните, тъй като ще се избегне нуждата от повтарящата се обработка на суровите данни.
- Разделяне на данни - тази стъпка също е изключително важна част от процеса по осигуряване на точността на модела за машинно обучение. Разделянето на данните спомага да се избегне преобучаването, позволява моделът да работи по-добре с нови данни и позволява да бъде оценено представянето му. В разширената архитектура би било подходящо да се дава възможност за разделяне на наборите от данни по различни критерии и методи, например hold-out и k-fold.
- Обогатяване на данни - обогатяването на данни е особено ценна стъпка, особено когато количеството или разнообразието от данни е ограничено. Идеята е изкуствено да се създадат данни, които да подобрят процеса по обучение на модела, използвайки данните, които вече имаме на разположение. Могат да бъдат използвани различни методи в зависимост от вида на данните.
- Намаляване на размерността - намаляването на размерността спомага за справяне със сложността на многомерните данни като намали броя характеристики, запазвайки основната информация. В разширената архитектура могат да бъдат използвани различни методи като метод на главните компоненти (PCA), който намалява броя дименсии като трансформира потенциално зависими една от друга променливи в по-малка група променливи, наречена главни компоненти. Това би направило наборите от данни по-лесни за съхранение и работа, без да бъде влошено представянето на модела.

Заключение

Складовете от данни играят решаваща роля в управлението и съхранението на данни за анализ, като предлагат добри възможности за първоначална обработка на данни. Чрез разширяването на архитектурата на традиционните складове данни, така че да включват допълнителни стъпки за подготовка на данни за машинно обучение, може да се автоматизират някои от процесите и да се улесни работата по подготовка на данни, която обикновено е скъп и времеемък процес. Добавянето на стъпки като подбор на характеристики, разделяне на данни, обогатяване на данни и намаляване на размерността могат да бъдат постигнати по-точни модели за машинно обучение без да се използват допълнителни ресурси.

References

1. Whang SE, Roh Y, Song H, Lee JG. Data collection and quality challenges in deep learning: a data-centric AI perspective. The VLDB Journal. 2023;32(4). doi:https://doi.org/10.1007/s00778-022-00775-9
2. Stonebraker M, Rezig E. Machine Learning and Big Data: What Is Important? Accessed October 29, 2024. http://sites.computer.org/debull/A19dec/p3.pdf
3. Roh Y, Heo G, Whang SE. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering. 2019;33(4):1-1.
4. What is Data Collection in Machine Learning and its Types? | SmartOne.ai. Smartone.ai. Published 2024. Accessed October 30, 2024. https://smartone.ai/blog/what-is-data-collection-in-machine-learning-and-its-types/
5. Krishnan S, Franklin MJ, Goldberg K, Wang J, Wu E. ActiveClean. Proceedings of the 2016 International Conference on Management of Data. Published online June 26, 2016. doi:https://doi.org/10.1145/2882903.2899409
6. Chu X. Data Cleaning.; 2019. https://bpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/b/1653/files/2020/10/data-cleaning-book-chapter.pdf
7. What is Data Transformation | Glossary. Hpe.com. Published 2024. https://www.hpe.com/emea_europe/en/what-is/data-transformation.html
8. What is Data Labeling? - Data Labeling Explained - AWS. Amazon Web Services, Inc. https://aws.amazon.com/what-is/data-labeling/
9. Li J, Cheng K, Wang S, et al. Feature Selection. ACM Computing Surveys. 2018;50(6):1-45. doi:https://doi.org/10.1145/3136625
10. Wang R, Tang K. Feature selection for MAUC-oriented classification systems. *Neurocomputing*. 2012;89:39-54. doi:https://doi.org/10.1016/j.neucom.2012.01.013
11. Reitermanová, Z. (n.d.). *Data Splitting*. https://physics.mff.cuni.cz/wds/proc/pdf10/WDS10_105_i1_Reitermanova.pdf
12. *What is Data Augmentation? - Data Augmentation Techniques Explained - AWS*. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/data-augmentation/
13. IBM. What is principal component analysis? | IBM. www.ibm.com. Published December 8, 2023. https://www.ibm.com/topics/principal-component-analysis