sciendo

Műszaki Tudományos Közlemények vol. 13. (2020) 148–151.
DOI: Magyar: https://doi.org/10.33895/mtk-2020.13.27
Angol: https://doi.org/10.33894/mtk-2020.13.27

EME MTSZ MTK

GÉPI TANULÁSBAN ALKALMAZOTT DIMENZIÓSZÁM-CSÖKKENTŐ MÓDSZEREK

DIMENSIONALITY REDUCTION METHODS USED IN MACHINE LEARNING

Muhi Kristóf,[1] Johanyák Zsolt Csaba[2]

Neumann János Egyetem, GAMF Műszaki és Informatikai Kar, Kecskemét, Magyarország

[1] *muhi.kristof@gamf.uni-neumann.hu*

[2] *johanyak.csaba@gamf.uni-neumann.hu*

Abstract

In most cases, a dataset obtained through observation, measurement, etc. cannot be directly used for the training of a machine learning based system due to the unavoidable existence of missing data, inconsistencies and high dimensional feature space. Additionally, the individual features can contain quite different data types and ranges. For this reason, a data preprocessing step is nearly always necessary before the data can be used. This paper gives a short review of the typical methods applicable in the preprocessing and dimensionality reduction of raw data.

Keywords*: machine learning, dimension reduction, data processing, big data.*

Összefoglalás

A megfigyelésekből szerzett nagy mennyiségű adat legtöbb esetben nem alkalmazható azonnal közvetlenül egy számítási intelligencián alapuló modell tanítására, mivel a gyakorlatban szinte elkerülhetetlen az, hogy hibás, inkonzisztens vagy hiányos adatokat tartalmazzon az adathalmaz. Emellett a különböző jellemzők értékei nagyon eltérő típusúak vagy nagyságrendűek lehetnek, ami különböző átalakításokat tehet szükségessé. A cikkben áttekintjük az adatfeldolgozás tipikus lépéseit.

Kulcsszavak*: gépi tanulás, dimenzió csökkentés, adatfeldolgozás.*

1. Adatok előfeldolgozása

1.1. Adatok tisztítása

A mintaadatok alapján történő tanításnál a rendelkezésre álló adathalmaz ún. minták sokasága. Minden mintát egy rekord vagy vektor ír le, azaz egy értéksor, amely különböző jellemzők értékeit tartalmazza az adott minta esetében. Például a *KDD Cup* 99 [1] adatbázis esetében 41 megfigyelt jellemző van, és 4 898 431 adatrekord.

Az adattisztítás során kiszűrjük azokat a rekordokat, ahol valamely tulajdonságnál nem a megengedett típusú érték áll (pl. hibás protokollazonosító). Emellett az adattisztítás feladata lehet az ismétlődések eltávolítása, mivel a duplán tárolt adatok hibás, félrevezető statisztikákat eredményezhetnek.

1.2. Hiányzó adatok

Az adatgyűjtés eredményeként előálló adathalmaz egyes rekordjai hiányosak lehetnek (pl. egy vagy több jellemző értéke nem lett rögzítve). Ilyen esetekben az alábbi stratégiák közül választhatunk.

1.2.1. Az érintett rekordok (vektorok) elhagyása

Ez a legegyszerűbb megoldás. Ha egy többezres mintahalmazban néhány rekord hiányos (pl. 1% alatt van az érintett rekordok száma), általában

különösebb kockázat nélkül elhagyhatjuk az érintett rekordokat. Mielőtt ezen opció mellett döntenénk, érdemes azt is megfontolni, hogy nincs-e jelzésértéke az adott attribútum értékhiányának, pl. hálózati forgalomnál nem lehetséges-e, hogy pont egy támadás eredményeképpen nem sikerült rögzíteni az érintett adatot.

1.2.2. A hiányzó adatok helyettesítése az adott attribútum átlagértékével

Ez egy egyszerű megoldás, amely viszonylag könnyen megvalósítható, azonban cserébe nagy lehet annak a kockázata, hogy a kapott adathalmazból teljesen hibás következtetéseket vonunk le.

1.2.3. A hiányzó adatok pótlása egyesével, emberi munkával

A megoldás néhány tíz érintett rekord esetében még reális lehet, azonban nagyon költséges és időigényes. Megvalósításának előfeltétele az, hogy valamilyen a priori ismerettel rendelkezzünk a vizsgálat tárgyára vonatkozóan.

1.2.4. A hiányzó adatok pótlása regressziós/ interpolációs technika segítségével

A módszer olyan esetben alkalmazható, ha valamilyen szabályosság (pl. lineáris változás) figyelhető meg az adott attribútum egymás utáni rekordokban megfigyelt értékei között vagy az érintett attribútum ún. „függő jellemző", azaz aktuális értéke valamely más attribútum vagy attribútumok értékeiből következik. Amennyiben sikerül egy függő jellemzőt azonosítani, akkor az esetek többségében nem érdemes túl sok energiát áldozni a jellemző hiányzó értékeinek pótlására, ugyanis a dimenziószám-csökkentés során ettől az attribútumtól úgyis megfogunk válni.

1.3. Függő attribútumok

Függőnek nevezünk egy attribútumot akkor, ha értéke egyértelműen származtatható egy vagy több másik jellemző értékéből. A függő attribútum értéke egy redundáns adat, amely csak növeli a kezelendő adatmennyiséget, és ezáltal a számítások idő- és tárigényét, azaz költségét. A dimenziószám-csökkentés során egyebek között a függő attribútumok kiszűrésére is törekszünk.

1.4. Átalakítás numerikus alakká

A minták hasonlóságának értékeléséhez és különböző statisztikák számításához numerikus adatokra van szükségünk. Abban az esetben, ha a címkék/kategóriák/különböző szöveges adatok értékeinél egyértelmű a sorrend és az egymástól való távolság (pl. egyenletes), akkor a feladat könnyen megoldható. Például tegyük fel, hogy három lehetséges címkeértékünk van. Ezek a kicsi, közepes és a nagy. A sorrend egyértelmű, és feltételezhetjük, hogy az egymást követő értékek azonos távolságra vannak. Ilyenkor a három címke előfordulásait 1, 2, és 3-as értékekre cseréljük.

1.5. Számosságcsökkentés

Adatok alapján történő modellépítésnél gyakori probléma, hogy nagyon nagy mennyiségű adattal kell dolgoznunk. A nagy mennyiségű adat rendelkezésre állása hasznos, viszont túlzottan megnövelheti a modell felépítéséhez szükséges időt, ezért sok esetben arra kell törekednünk, hogy egyfajta optimumpontra állítsuk be a felhasznált adatmennyiséget úgy, hogy a lehető legkisebb rekordszám mellett a lehető legtöbb információt őrizzük meg az eredeti adathalmazból, azaz a kiválasztott részhalmaz legyen reprezentatív [2].

2. Dimenziószám-csökkentés

A modellfelállítás során a rendelkezésre álló adatrekordok gyakran nagyszámú jellemzőt tartalmaznak, ami részben arra is visszavezethető, hogy az adatgyűjtés megtervezésekor még nem rendelkeztek elegendő információval a megfigyelt jelenségre vonatkozóan, így minden elképzelhető, rögzíthető adatot gyűjtenek annak érdekében, hogy minél teljesebb képet kapjunk, és nehogy valami fontos kimaradjon.

A sok jellemző azonban nagy memória- és számításigényt is jelent, ami akár megoldhatatlanná is teheti a feladatot. Ezért törekednünk kell arra, hogy a feladat dimenzionalitását csökkentsük. A dimenziószám-csökkentő módszerek célja az, hogy az eredetileg m dimenziós adatpontokat úgy képezzük le egy k dimenziós térbe (ahol $k < m$), hogy [3]

- minél jobban megőrizzük a pontok osztályozhatóságát, azaz az egy osztályba/kategóriába tartozó pontok minél jobban elkülönüljenek a többi osztályba/kategóriába tartozó pontoktól második;
- minimális legyen az információ veszteség.

Számos módszer kínálkozik a feladatra. Elsőként nézzük meg azokat, amelyek nem járnak információvesztéssel, azaz a redundáns adatok kiszűrését célozzák meg.

2.1. Információvesztés nélküli dimenzió-szám-csökkentés

Azoknál az adathalmazoknál, amelyek a „mindent gyűjtünk" elv alapján keletkeztek, könnyen előfordulhat, hogy egy jellemző értéke minden rekordban azonos. Ilyenkor ezt a jellemzőt (vagy oszlopot, ha egy táblázatban/mátrixban gondolkodunk) minden további nélkül eltávolíthatjuk adatbázisunkból. Bár ez egy triviális megoldásnak tűnik, de a gyakorlatban könnyen előfordulhat (ld. az NLS-KDD [4] adatbázis 20%-os mintanagyságú tanító adatbázisa).

2.2. Dimenziószám-csökkentés főkomponens elemzéssel

A legtöbb gyakorlati feladatnál kismértékű információvesztés elfogadható kockázatot jelent, ha cserébe jelentős mértékű dimenziószám-csökkenést lehet elérni. Tekintsünk a rendelkezésünkre álló mintaadatokra mint pontokra egy sokdimenziós térben. Ekkor a mintaadatrekord minden egyes eleme a pont egy koordinátáját jelenti ebben a térben. Számos feladatnál ezek a pontok nem véletlenszerűen helyezkednek el a térben, hanem valamilyen szabályosságot követnek, vagy a pontok változékonysága nem minden irányban egyforma.

Például az 1. ábrán a pontok közelítőleg egy egyenes mentén helyezkednek el. Ha felrajzoljuk ezt a képzeletbeli egyenest egy koordinátatengelyként (c_1 az 1. ábrán) és a rá merőleges koordinátatengelyt is berajzoljuk, akkor megfigyelhetjük, hogy a pontok nagy változékonyságot mutatnak a c_1 tengely mentén, és viszonylag kicsi a változékonyságuk az arra merőleges c_2 irányban.

Amennyiben a pontok merőleges vetületeit képezzük a c_1 tengelyre, akkor egy olyan pontsort kapunk, amely a tengelyen helyezkedik el (piros pontok a 2. ábrán). Ezek a pontok csak kismértékben térnek el az eredeti pontoktól, viszont helyzetük egyetlen koordinátával leírható a c_1 tengely mentén. Amennyiben a továbbiakban ezekkel a pontokkal dolgozunk az eredetiek helyett, akkor az eredetileg kétdimenziós adathalmaz dimenziószámát eggyel csökkentettük kismértékű adatvesztéssel.

A főkomponens-elemzés célja ezen tengelyek beazonosítása, majd a pontok koordinátáinak meghatározása az új koordinátarendszerben egy lineáris transzformáció segítségével.

2.3. Információnyereség-alapú dimenzió-szám-csökkentés

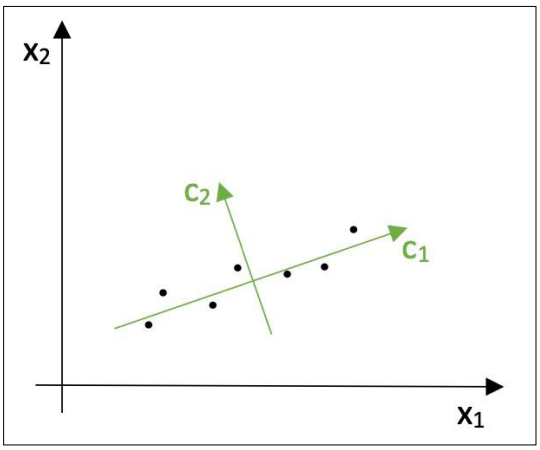
Az információnyereség- (Information Gain – IG) alapú dimenziószám-csökkentés módszerének alapgondolata a döntési fák elméletéhez (pl. ID 3, C4.5, C5.0 algoritmusok [5]) kapcsolódik, ahol jellemzőkként haladva a kiválasztott jellemző értékei szerint minden döntésnél részhalmazokra bontja a kezdeti halmazt. Az algoritmus alapja az entrópiaszámítás. Az S halmaz entrópiája

$$E(S) = -\sum_{k=1}^{N} p_k \log_2 p_k \qquad (1)$$

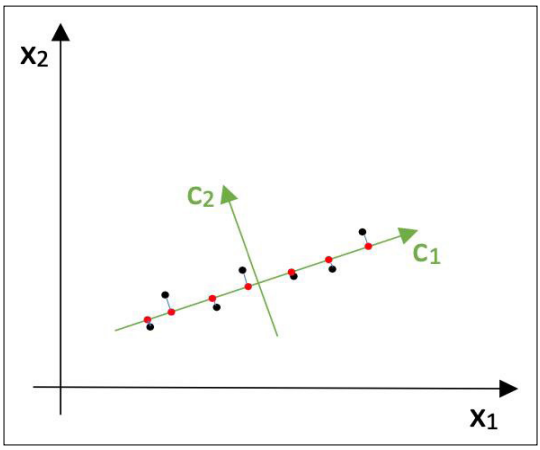
a halmaz "szennyezettségét" (inhomogenitását), azaz változékonyságát jellemzi. Itt N a halmazban előforduló értékek száma és p_k az egyes értékek relatív gyakorisága, amit a

$$p_k = \frac{n_k}{n} \qquad (2)$$

képlettel számítunk, ahol n az összes halmazelem, n_k a k. címkével rendelkező halmazelemek száma.



1. ábra. *Ponthalmaz új koordináta tengelyekkel*



2. ábra. *Vetítés a c_1 tengelyre*

2.4. Véletlen projekciók (Random Projection)

Az eljárás a pontokat az eredeti m dimenziós térből egy alacsonyabb r dimenziószámú térbe vetíti véletlenszerű lineáris projekciót alkalmazva. A vetítés úgy történik, hogy egy véletlen számokból előállított mátrixxal megszorozzuk a jellemzők mátrixát.

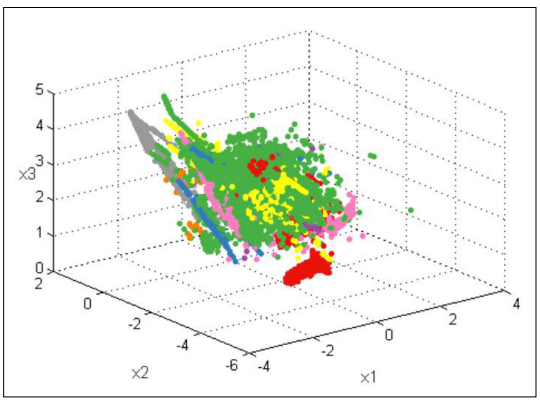
$T = X \cdot RP$, ahol $T \in \mathbb{R}^{nxr}, X \in \mathbb{R}^{nxm}, RP \in \mathbb{R}^{mxr}$ és RP minden sora egység nagyságú vektor

$\| RP_i \|^2 = 1, j = 1..m.$

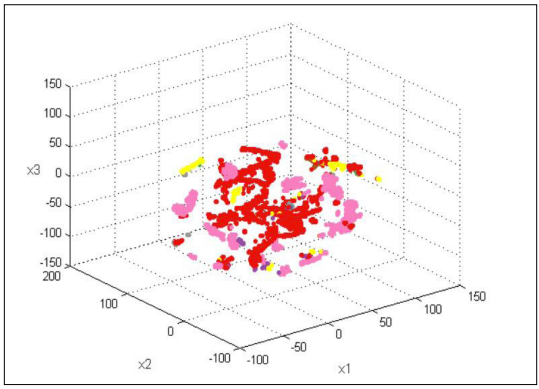
A véletlen számok Gauss-eloszlását követik. A módszer jól megőrzi a pontok közötti távolságot, és számításigénye kisebb a PCA-nál. A transzformáció minősége a pontok számától és r értékétől függ. A **3. ábra** az NSL-KDD 20%-os tanító adathalmaz egy 3D-s véletlen projekcióját mutatja be. A pontok színei a forgalomtípusokra utalnak.

2.5. T-eloszlású sztochasztikus szomszéd beágyazás (t-Distributed Stochastic Neighbor Embedding)

A t-DSNE eljárás a dimenziószám-csökkentés során arra törekszik, hogy a hasonló pontokat egymáshoz közel tartsa, míg az eltérőket egymástól távol **[3]**. Az eredeti pontokat úgy modellezi, mintha azok normál eloszlásból származnának, a beágyazott pontokat pedig úgy, mintha azok egy Student(t) eloszlásból származnának. Legtöbbször pontcsoportok (klaszterek) két- vagy háromdimenziós vizualizációjára használják. Az eljárás hátránya, hogy eredményeképpen nem keletkezik egy mátrix vagy képlet, amely lehetővé tenné azt, hogy további (pl. teszt, validációs stb.) adatokat ugyanabba a térbe transzformáljunk az eredeti m dimenziós térből. A **4. ábra** bemutatja a t-DSNE-transzformáció eredményét a NSL-KDD



3. ábra. *Az NSL-KDD 20%-os tanító adathalmaz egy véletlen projekciója a háromdimenziós térbe*



4. ábra. *Az NSL-KDD 20%-os tanító adathalmazból véletlenszerűen kiválasztott 10 000 minta t-DSNE-transzformációja a háromdimenziós térbe Chebysev-távolságokat alkalmazva*

20%-os tanító adathalmazból véletlenszerűen kiválasztott 10 000 minta esetére.

Összegzés

A gépi tanulás során felhasználni kívánt nyers mA gépi tanulás során felhasználni kívánt nyers mintaadatok legtöbbször egy előfeldolgozási lépésen kell, hogy átessenek a tényleges felhasználást megelőzően. Ez a lépés magában foglalja a hiányzó adatok problémájának kezelését, a számosság és a dimenzionalitás csökkentését. Cikkünkben röviden áttekintettük az ilyen esetekben leggyakrabban alkalmazott megoldásokat.

Köszönet

Köszönettel tartozunk a kutatás támogatásáért, amely az EFOP-3.6.1-16-2016-00006 „A kutatási potenciál fejlesztése és bővítése a Neumann János Egyetemen” pályázat keretében valósult meg. A projekt a Magyar Állam és az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával, a Széchenyi 2020 program keretében valósul meg.

Szakirodalmi hivatkozások

[1] KDD Cup 1999 Data http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[2] Ősz R., Holik I.: *Pedagógiai kutatásmódszertan*, Óbudai Egyetem, 2015.

[3] Géron A.: *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems*, O’Reilly, 2019. ISBN 978-1-492-03264-9

[4] NSL-KDD dataset, https://www.unb.ca/cic/datasets/nsl.html

[5] Mitchell T. M.: *Machine learning.* McGraw-Hill Science/Engineering/Math, 1997.