BUDAPEST MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM

GÉPÉSZMÉRNÖKI KAR

GYÁRTÁSTUDOMÁNY ÉS -TECHNOLÓGIA TANSZÉK

Félvezetők gyártási hibájának automatizált felismerése tömeges szenzor-adatok alapján

Tudományos Diákköri Konferencia Budapest, 2021

Készítette:

Pitlik Marcell

mechatronikai mérnök MSc hallgató

Konzulens:

Ibriksz Norbert

doktorandusz

NYILATKOZAT AZ ÖNÁLLÓ MUNKÁRÓL

Alulírott, Pitlik Marcell (ZUBYPT), a Budapesti Műszaki és Gazdaságtudományi Egyetem hallgatója, büntetőjogi és fegyelmi felelősségem tudatában kijelentem és sajátkezű aláírásommal igazolom, hogy ezt a TDK dolgozatot meg nem engedett segítség nélkül, saját magam készítettem, és a TDK dolgozatomban csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Budapest, 2021. november 7.

Pitlik Marcell

.

Tartalomjegyzék

Absztrakt 3

Abstract 4

1. Bevezetés 5

1.1. Miért a SECOM-adatbázist került kiválasztásra? 5

1.2. Modellezési kihívások 5

2. Szakirodalmi alapozás 6

2.1. Adat-centrikus szemlélettel 6

2.2. Modell-centrikus szemlélettel 6

2.3. 80:20 szcenárió 7

2.4. 70:30 szcenárió 12

2.5. Általános észrevételek 14

3. Adatvagyon – saját elemzésekhez 16

3.1. Módszertani alapvetések 16

3.2. MCM-hibridmodell 17

3.2.1. Alapszint 17

3.2.2. Zárószint 17

3.3. Y0-alapú hibrid modell 17

3.4. Objektumokat lépcsőzetesen a tanulási mintából kizáró modell-láncok 18

4. Eredmények 20

5. Speciális kihívás – anti-diszkriminatív modellezés 23

6. Konklúziók 25

7. Jövőkép: Optimalizálás rétegei 26

8. Felhasznált források 27

9. Mellékletek 28

ABSZTRAKT

Az önmagában is automatizált tömeggyártás kapcsán a gyártási folyamat szenzoros megfigyelése magától értetődően képes big-data-jellegű erőtereket felkínálni a félvezetők esetén is (vö. SECOM-adatbázis). Az elemzési kérdés is látszólag egyértelműen adott: olyan modelleket kell fejleszteni, melyek alapján a tanulási adathalmazban eleve alacsony arányszámban fellelhető hibás termékek a szenzoradatok alapján úgy ismerhetők fel, hogy a levezetett összefüggés az (éles) tesztadatokra is racionális pontossággal érvényes lesz. A probléma azért látszólagos, mert ennyiből még közel sem triviális, mennyi hibátlan félvezető tévesen hibásként való klasszifikálása és/vagy mennyi hibás, de hibátlannak vélt félvezető piacra dobása milyen hasznossági hatásmechanizmusokkal bír? Más megfogalmazásban: mi a hatás-szintű ekvivalencia (értékazonosság) eltérő jóságrétegekkel rendelkező modellek esetén? Még általánosabban: melyik modell a legjobb? Ezen elméleti kérdés is tételesen értelmezésre kerül az anti-diszkriminatív modellezésre támaszkodva, ahol a centrális kihívás annak bizonyítása – lehet-e minden kontingencia-táblázat másként egyformán értékes – különösen a teoretikus alapvetéshez, mint potenciális normához képest, ahol a rel. alacsony előfordulási gyakoriságú hibás félvezetőkre vonatkozó modellbecslés nem létezik. A lépcsős függvény-alapú klasszifikáló modell-láncok, melyek itt és most bemutatásra kerülnek képesek arra, hogy alapvetően a hibátlanságot, mint olyat egyetlen egy becslési értékként legyenek képesek megjeleníteni, ahol tehát a tanulási adatok esetén soha nincs hibás termék. A hibátlanságért felelős modell-outputérték a tesztadatok esetén azonban sajnos nem kizárt, hogy továbbra is tartalmaz hibás termékeket, s hasonlóképpen: a speciális output-értéken kívüli minden más becslési érték zömmel hibás termékre utal, de itt sem kizárólagosan. Ezt a képességet egyrészt sűrített, nem-monoton lépcsős függvényekkel és súlyozott output-értékekkel lehet elérni, ahol a sűrítés és a súlyozás maga is optimalizálható. Emellett az egyetlen karakteres outputértékkel jellemezhető objektumok kizárása a tanulási mintákból sokkal gyorsabb tanulást és jobb teszteredményeket tesz lehetővé. Maga az irányítatlan lépcsős függvény, ill. a lépcsős függvény-típusok és rétegek hibridizációja modell-orientált beavatkozás. Így a modell-centrikus és adat-centrikus mesterséges intelligencia-fejlesztés kapcsán egyszerre kerülnek innovációk bemutatásra.

ABSTRACT

The automated production of semiconductors can be observed with quasi unlimited sensors and so, the big-data can be ensured for classification projects (s. SECOM-database). The question is seemingly trivial: how it is possible to derive a classification model with the less error (like false negative and false positive). The ratio of false positive and false negative objects has however no trivial direction: the equivalences are economical question especially in unbalanced cases like in SECOM-database. The bast model can not be defined in a simple way. The anti-discriminative modelling approach makes possible to derive an aggregated goodness index where the theoretical scenario of the massive-unbalanced training sets prescribe in a quasi norm-like situation, that no fails may be estimated only passes. The focused solutions (the staircase functions) are capable to derive a constant output value for the majority class. The challenge is whether this specific estimation value becomes a pure rule, or the test cases makes it dirty. The presented models will involve chains of staircase functions with innovative features like reducing processed pixels, exclusion objects, hybridization of staircase layers and/or types.

1. BEVEZETÉS

1.1. Miért a SECOM-adatbázist került kiválasztásra?

A tanulmány, ill. a mögöttes vizsgálatok apropóját egy egyetemi projekt adta. A SECOM-adatbázis [1] beazonosítása, letölthetősége és az erre alapozó publikusan elérhető cikkek számossága alapján az adatvagyonként a SECOM-adatbázis került elfogadásra.

1.2. Modellezési kihívások

Az elemzési kérdés látszólag egyértelműen adott: olyan modelleket kell fejleszteni, melyek alapján a tanulási adathalmazban eleve alacsony arányszámban fellelhető hibás termékek a szenzoradatok alapján úgy ismerhetők fel, hogy a levezetett összefüggés az (éles) tesztadatokra is racionális pontossággal érvényes lesz. A probléma azért látszólagos, mert ennyiből még közel sem triviális, mennyi hibátlan félvezető tévesen hibásként való klasszifikálása és/vagy mennyi hibás, de hibátlannak vélt félvezető piacra dobása milyen hasznossági hatásmechanizmusokkal bír? Más megfogalmazásban: mi a hatás-szintű ekvivalencia (értékazonosság) eltérő jóságrétegekkel rendelkező modellek esetén? Még általánosabban: melyik modell a legjobb? Ezen elméleti kérdés is tételesen értelmezésre kerül az anti-diszkriminatív modellezésre támaszkodva, ahol a centrális kihívás annak bizonyítása – lehet-e minden kontingencia-táblázat másként egyformán értékes – különösen a teoretikus alapvetéshez, mint potenciális normához képest, ahol a rel. alacsony előfordulási gyakoriságú hibás félvezetőkre vonatkozó modellbecslés nem létezik.

2. SZAKIRODALMI ALAPOZÁS

A nagy számosságú SECOM-érintettségű szakirodalmi utalásból quasi véletlenszerűen 2 publikáció lett kiragadva. (17200 találat a ’SECOM database semiconductor’ a google kereső motorjával)

1. A továbbiakban 80:20-as esetnek nevezett tétel [2]
2. S a továbbiakban 70:30-as esetnek nevezett tétel [3]

A szakirodalmi helyzetértelmezés kapcsán kiemelendő karakterisztikák (zárójelben a saját specifikus megközelítésekkel):

2.1. Adat-centrikus szemlélettel

- Ha egyáltalán, akkor hogyan kerülhet az attribútum-készlet redukálásra?
- Ha egyáltalán, akkor hogyan kerülhet az objektum-készlet átalakításra?
- (Hogyan hatnak az adat-szintű konverziók a klasszifikálás eredményességére?)
  - (az output-arányok és/vagy sorrendek befolyásolása kapcsán)
  - (a pixelesség (vagyis a folytonos inputok egyre kevesebb diszkrét halmazba sorolása kapcsán)

2.2. Modell-centrikus szemlélettel

- Milyen adatfeldolgozási (tudás-reprezentációs) technikák kerülhetnek bevonásra?
- (Hogyan hat a modell-hibridizáció a klasszifikációk eredményességére?)
  - (a több-rétegű azonos típusú modellek kapcsán)
  - (a több rétegű vegyes típusú modellek kapcsán)

Az adat-centrikusság és a modell-centrikusság fogalompárjának forrása Andrew Ng videója [9] volt:

A szakirodalom elemzése rámutatott több kritikus minőségbiztosítási pontra is:

- A true-positive fogalom kettős értelmezésének lehetőségére és tényére (már egy cikken belül is).
- A reprodukálhatóságot masszívan korlátozó (esetlegesen kizáró) közléspontosság sorozatos tetten érthetőségére (vö. eredmények: csak tanulás, csak teszt, mindösszesen, ill. tanulási minták pontos rekordszáma, ill. a modellparaméterek hiánya).

A saját vizsgálatok kapcsán tehát semmilyen formában nem volt elvárható a tételes összehasonlíthatóság garantálása, de cikk üzenetéhez erre nem is volt szükség.

A cikk célja tehát annak demonstrálása, hogy quasi véletlenszerű modellezési paraméterek kiválasztásával a szinte végtelen paramétertérben, ami a lépcsős függvények kapcsán a szerző által feltárása került, a klasszifikáció eredményessége már norma-szerű egy objektív modellértékelési eljárás alapján.

A norma-szerűség demonstrálni tudása elegendő a potenciál jelzéséhez minden olyan esetben, amikor több, mint 90%-a a rendelkezésre álló attribútumoknak nem került még egyáltalán felhasználásra, de a hibridizáció hasznának (vagyis a további attribútumok integrálhatóságának) bizonyítékai máris adottak.

2.3. 80:20 szcenário

Releváns részletek a publikációból: [2]

„The removal of the constant values deals with the elimination of those features from the dataset that have constant values for all the samples from the training data. After the removal of 116 features which consist of constant values, the number of the features that will be considered in the next steps becomes 451”

„Boruta Algorithm identifies the relevant variables by using a classification algorithm which is wrapped around the Random Forest classification algorithm. The package Boruta is implemented in R, a language which is used widely in statistics and data mining.”

„From the total of 451 features, 22 were confirmed as important while 428 were confirmed as unimportant or potentially important. The feature which does not appear in the results (451−22−428 = 1) is the label.”

Ha a rendelkezésre álló adatvagyon (attribútum-szám) esetén csak ennek töredékét (pl. 10%-át) választja ki tehát valaki, akkor a kanonizált megközelítések értelmében nem kizárt, hogy a töredékes információmennyiség elég lehet sikeres modellek alkotásához. Ennek a megközelítésnek azonban ellentmondani látszik az a tapasztalat, miszerint megfelelően flexibilis adatértelmező módszertanok mellett a minden-mindennel-összefügg-elv alapján nincs értékes és értéktelen attribútum! (vö.: parciális korreláció [4])

TABLE II
COMPARISON OF THE FEATURES SELECTION TECHNIQUES

Features Selection Algorithm	Number of Selected Features	Characteristics
Boruta	22	by default it uses the Random Forest (RF) algorithm
MARS	10	it is a form of regression analysis
PCA	111	it uses orthogonal linear transformations

1. ábra: Attribútum kiválasztási technikák összehasonlítása forrás: [2]

TABLE VI
FEATURES SELECTION

Approach	Number of Features
Worst Case	450
Boruta	22
MARS	10
PCA	111

2. ábra: Attribútum kiválasztás forrás: [2]

Az attribútumok priorizálása nem szükséges elsődlegesen a magas klasszifikáció sikerhez, amennyiben a priorizálást maga a modell-centrikus megközelítések is képesek elvégezni.

„The training data is used as input for three Classification Algorithms which are briefly described below. The Logistic Regression (LR) algorithm was used in at least three research papers that analysed the SECOM dataset. It is one of the simplest classification algorithms. The dependent variable is binary, and it can be either 1 or 0. The LR applied in the experimental results section uses the Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) optimization algorithm in order to use only a limited amount of the computer memory. We have used LR as a reference algorithm to evaluate the performance of the other two classification algorithms tried in this paper, the Random Forest (RF) and the Gradient Boosted Trees (GBT)which to the best of our knowledge were not applied before on the SECOM dataset."

TABLE V
CONFUSION MATRIX

	Actual Class = 1 (Fail)	Actual Class = -1 (Pass)
Prediction Class = 1 (Fail)	True Positive (TP)	False Positive (FP)
Prediction Class = -1 (Pass)	False Negative (FN)	True Negative (TN)

3. ábra: Kontingencia táblázat forrás: [2]

A kontingencia-táblázat (confusion matrix) mind a négy pozíciója és ezek aránya egyszerre fontos. Így előáll a filozófiai probléma: mi is a hibátlan modell után a második legjobb? Azaz milyen típusú hibák ekvivalensek és miért?

Azonban a teoretikus benchmark mindenkor létezik: az a modell az alapszintje (normája) a kontingencia-értékek egymással való versenyének, ahol a becslések szerint csak jó félvezető van.

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4. ábra: Statisztikai mutatók definíciója forrás: [2]

A mutatószámok kapcsán nincs érdemi észrevétel. Ezek léte is azt erősíti, hogy az összes alapszám (TP, TN, FP, FN) mindennemű aránya releváns – s az arányok a relevánsak, mert így a mintaméretektől lehet függetlenedni.

Itt kell megemlíteni, hogy a TP és a TN fogalma félreérthetetlen, de az FP és az FN fogalma logikai szinten kettősséggel bír:

- FP(1) = tévesen pozitívnak jelzett, valójában negatív objektum
- FP(2) = tévesen klasszifikált pozitív objektum

Ez a logikai (hermeneutikai) kettősség a 30:70-es [3] publikációban mindkét rétegével meg is jelenik.

TABLE VII
SAMPLING METHOD

Sampling Method	Description
Unsampled	the number of samples is unchanged
Undersampled Majority Class	the number of samples of the majority class is reduced to 78
Oversampled Minority Class	the number of samples of the minority class is increased to 1176

5. ábra: Tanulási minták forrás: [2]

A tanulási minta mérete és összetétele a fentiek értelmében tetszőlegesen manipulálható és elvárt, hogy ennek érdemi hatása legyen a klasszifikációk sikerességére, de nem tudni előre adott feladat esetén, melyik tanulási minta méret és arány lesz a leginkább célravezető, hacsak a sokféle véletlen tanulási minta leíró adatai (Xi) alapján nem építünk a tanulás sikerességét becslő termelési függvényt és nem kezdünk el genetikai potenciál-alapon tanulási mintát optimalizálni. [5]

Fig. 7. Experimental Results Summary - Features Selection Perspective

6. ábra: Közölt eredmények I. forrás: [2]

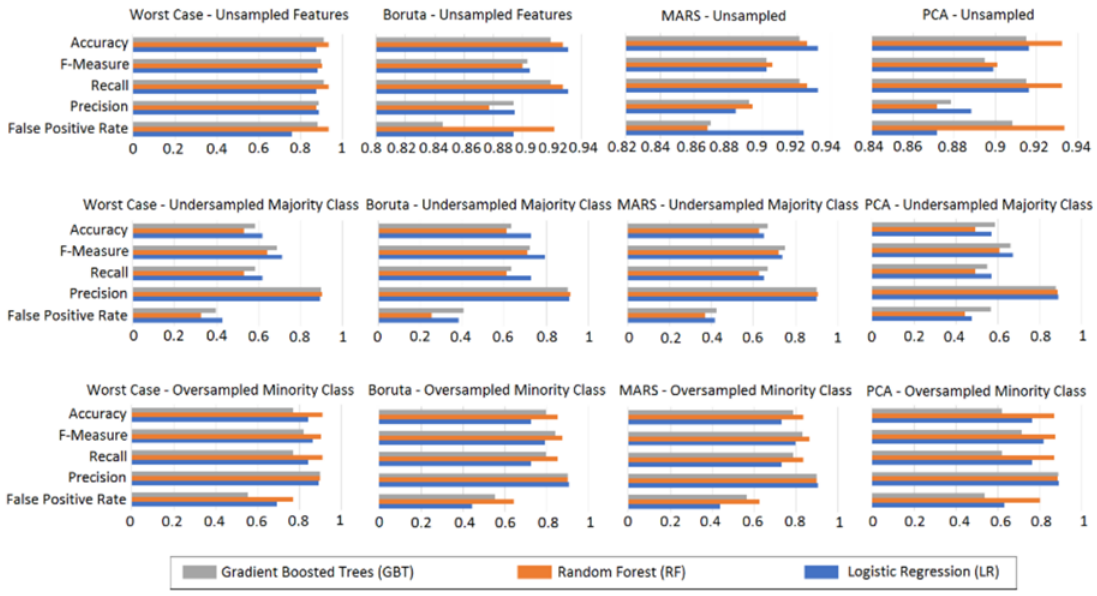


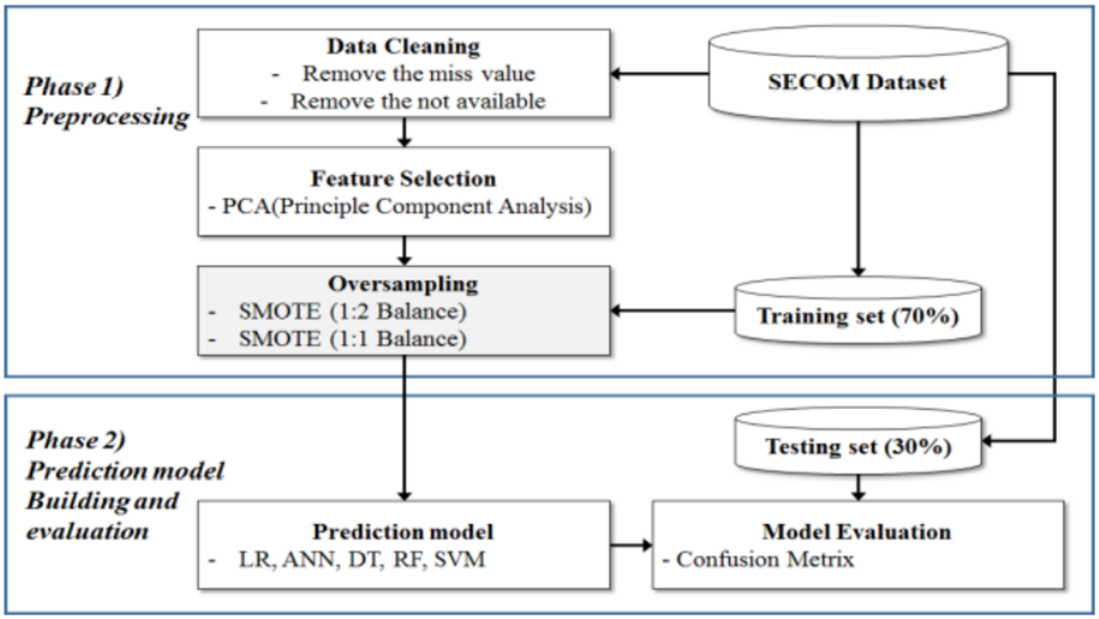
Fig. 8. Experimental Results Summary - Classification Algorithm Perspective

7. ábra: Közölt eredmények II. forrás: [2]

Ahhoz, hogy a fenti pontossági értékeket értelmezni lehessen, előbb tudni illik, vajon az 1567-es teljes objektumhalmazra létezik-e hibátlan tudás-reprezentáció (akár több is)?! Emellett ilyen eredményeket kötelező kellene, hogy legyen numerikus értékek formájában is átadni, már csak azért is, hogy a mindenkor versengő megoldási alternatívák (jelen esetben 3·3·4=36) anti-diszkriminatív rangsorolását el lehessen végezni, mert minden további versenyzőt csak így lehet érdemben pozícionálni. Erre a grafikus publikáció sajnos teljesen alkalmatlan, mert a 36 érték rangsora nem állapotható meg mutatószámról mutatószámra tetszőleges pontossággal a képek alapján.

2.4. 70:30 szcenário

Érdemi numerikus eredményt ez a publikáció tartalmazott, így a saját elemzések 80:20-as abszolút számai a 70:30-as arányokhoz lettek igazítva.



8. ábra: Modellezés lépései forrás: [3]

Az SVM rövidítés csak a 8. ábra: szerepel – a publikációban már sehol sem érhető tetten. Cserében az ábra nem utal a RUS-ra, amire viszont a szöveg több ponton is.

		Actual Class	
		Class=1 (Fail)	Class= -1 (Pass)
Prediction Class	Class=1 (Fail)	True Positive	False Positive
	Class= -1 (Pass	False Negative	True Negative

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(FP + TN)$$

$$Accuracy = TP + TN/(TP + TN + FP + FN)$$

$$Precision = TP/(TP + FP)$$

$$F - measure = 2TP/(2TP + FP + FN)$$

9. ábra: Statisztikai mutatók definíciója forrás: [3]

A 80:20-as és a 70:30-as publikáció értékelési rendszere látszólag azonos: vagyis a 104 hibás félvezető tesztbe jutó részét (70:30 esetén: 30 darabot) elvileg a TP+FN összege kellene, hogy kiadja.

A TP+FP a becslés kapcsán előállt hibásnak vélt objektumok száma kellene, hogy legyen – ami alapvetően ritkán lehet konstans érték sok-sok modell esetén. Ezzel szemben:

Confusion matrix result

Sampling Method	Prediction Model	TP	FP	FN	TN
SMOTE 1:2	LR	14	16	61	377
	ANN	5	25	27	411
	DT	5	25	69	369
	RF	7	23	27	411
RUS 1:2	LR	17	13	107	331
	ANN	18	12	173	265
	DT	7	23	69	369
	RF	5	25	20	418

10. ábra: Közölt számszerű eredmények forrás: [3]

a TP+FP = minden sor (megoldási alternatíva) esetén félreérthetetlenül 30 darab, ami üti mindkét publikáció elméleti alapjait. De szerencsére semmilyen formában nem hat ki az anti-diszkriminatív értékelésre, mely ilyen értelemben context free (oszlop-fejléc-független), hiszen csak az számít, hogy minden T* minél nagyobb annál jobb és minden F* minél kisebb annál jobb értelmezést kapjon. Így a helyes származtatott mutatószámok (arányszámok) azok, ahol a számlálóban T*, a nevezőben F* található, mert ekkor az irány: minél nagyobb, annál jobb, vagy ennek inverze inverz iránnyal.

2.5. Általános észrevételek

A két publikáció tehát többféle modell-centrikus technikát említ: vö.

- Döntési fa (DT)
- Mesterséges neurális háló (ANN)
- Random Forest (RF)
- Logisztikus regresszió (LR)
- Gradient Boosted Trees (GBT)

A lépcsős függvényekre alapozva párhuzamosan és hibridizálva lehet

- Irányítatlan (döntési fa-jellegű) modelleket alkotni (vö. COCO-MCM)
- Irányított modelleket alkotni (COCO-STD, COCO-Y0), ahol a dupla-attribútumkészlettel dolgozó STD-modellek hidat képeznek az MCM felé…

Az objektumok számát befolyásoló technikák meta-szinten:

- Unsampled
- Oversampled
- Undersampled

Eljárás-szinten: pl.

- SMOTE
- RUS, …

Saját megközelítésként értelmeződik a következőkben: a homogén outputokat alkotó elő-modellek objektumhalmazai és ezek kizárása a további modellezésből.

Az attribútumok számát befolyásoló megoldások:

- PCA (mindkét helyen említve)
- MARS
- Boruta
- Worst Case, …
- Ill. egyéb statisztikai, adathiány-kezelési típusokra alapozó megoldások

A saját megoldás az attribútum szám redukálására nem módszertani, hanem egyenszilárdsági elvet követett: minden attribútum, mely nem volt megadva mind az 1567 objektum esetén, ki lett zárva ennek minden további értelmezése nélkül annak érdekében, hogy az adathiány-kezelés mibenlétének önálló módszertani problémája itt és most ne terhelje az eleve nem teljesen összehasonlítható paraméterhalmazokat.

A tanulási minta képzésnek két speciális (saját) megoldása az alábbi volt:

- A függő változó bináris értékeinek numerikus nagysága és ezek sorrendje egymáshoz képest, hiszen az under/over-sampling ezzel a paraméterrel is szimulálható.
- Az Xi értékek esetleges folytonos változók esetén is essenek felbontás-redukció alá, azaz eleve és tudatosan pixeles képből induljunk ki nagyfelbontású kép helyett annak érdekében, hogy a döntési fa-jellegű tudásreprezentáció szabályait erősítő objektumhalmazok kikényszerítésre kerüljenek (jelen esetben alapvetően azonos méretű halmazok formájában a rangsorszámmá konvertált inputok adott konstanssal való elosztása révén).

3. ADATVAGYON – SAJÁT ELEMZÉSEKHEZ

Az adatvagyon letölthető az alábbi URL-ről: [1]

Az esetek/termékek/objektumok (félvezetők) száma mindösszesen: 1567

A tanulás és a teszt aránya: 80%:20% (4:1) (1254+313=1567)

A teszt az időpecsét szerint rendezett alapsokaság utolsó 20%-a.

A hibás termékek száma/aránya a teljes sokaságban: 104 (ill. 104/1567=6.64%)

A hibás termékek száma/aránya a tanulásban: 87/1167=7.46%

A hibás termékek száma/arány a tesztben: 17/313=5.43%

A teszt/tanulás arány a hibás termékek esetén: 17/87 (kb. 1:4), ill. 17/104=16.34%

A context free módon értelmezett szenzor-attribútumok (Xi) száma: 591 – folytonosnak tekintett változók

Ebből a teljes alapsokaság esetén hiánytalan attribútumok (Xi) száma: 52 (52/591=8.79%)

A következmény-változó (Y): hibás (-1) vagy hibátlan (1) - bináris

A felhasznált tanulási OAM és teszt OAM paraméterei: 1254·(52+1), ill. 313·(52+1)

3.1. Módszertani alapvetések

A hasonlóságelemzés lépcsős függvényeit több, egymástól független módon lehet előállítani:

- Y0 (anti-diszkriminatív) modell esetén a lépcsők szigorúan monotonok, előzetesen megadott irány-preferenciákat követve
- STD (termelési függvény) modell esetén a lépcsők monotonok, előzetesen megadott irány-preferenciákat követve – kivéve dupla-attribútum-készlettel dolgozó modellek esetén, ahol az iránypreferenciák optimalizálása történik meg
- MCM (exploratív) modell esetén (vö. döntési fák) a lépcsőknek nincs iránya – a lépcsők értékének robosztusságát az egy (csökkentett számú) lépcsőfokba sorolt objektumok (pixelek) száma adja a minél nagyobb, annál robosztusabb elv alapján – de a modell plaszticitását a minél több a lépcsők száma annál rugalmasabb a modell elv vezérli

3.2. MCM-hibridmodell

3.2.1. ALAPSZINT

Modell-rétegek száma az 1. szinten: 5 (10+10+10+11+11=52 attribútum)

Sorszámozási sűrítés: 50, azaz a nyersen sorszámozott tanulási adatok sorszámértékének ötvened része + 1 az új (sűrített) sorszám (sorszámok maximális értéke – vö. pixel: 26)

Teszt-sorszámok: darabhatöbb()+1, vagyis a tanulási mintában hány darab adat nagyobb, mint egy-egy tesztadat

Lépcsőszám (tanulás és teszt): 1, …, 26

Y-variánsok (jó:rossz):

- 1000:9000 (1167 vs. 87, ill. 303 vs. 87, ill. 57:87),
- 99000:9000 (87:57),
- 1000:99000 (57:87)

Futtatási környezet: MYX-FREE COCO MCM [6]

3.2.2. ZÁRÓSZINT

A záró-modell objektumszáma azonos, mint az alapszint esetén a tanulásban is és a tesztben is.

A zárómodell attribútumai (Xi): 5+1, vagyis az ötalapszintű becslés értéke és ezek objektumonkénti szórása/átlaga az alapszintre jellemző sűrítés (vö. 50, azaz max. 26 lépcső) mellett.

Y-attribútum azonos az alapszint adataival a tanulásban és a tesztben is.

A záró-modell futtatási környezete azonos a tanuláséval.

3.3. Y0-alapú hibrid modell

A teljes adatvagyon kezelése mellett készült 87+87 elemű tanulási minta is (a 80:20-as forrást követve), ahol a 87 jó félvezető véletlenszerűen lett kiválasztva, melynek tanulási paraméterei:

- 3 réteg = [(1+1)+1]+1 modell

- direkt és inverz (1+1) alapmodellek (Y0) alapján
- ezek karakterisztikáiból álló attribútumokra MCM-modell épült
- mely output-skáláján alsósávként megjelentek tömegesen a hibátlan félvezetők
- középső részén tömegesen a hibásak, ill.
- felső részén 40+11 arányban a jók és a rosszak szinte homogén elegye
- de a felső tartomány egy önálló szeparátor-modellel (MCM) hibátlanul felbonható volt és a legsikeresebb teszteredmények ennek a szeparátor-modellnek a termékei (ami megalapozta a 1254->390->144-es modell-lánc objektumkizáró stratégiáját)

- 87+87 rekord
- 52 attribútum /
- 9&18 pixel
- Y0&MCM ->0 tanulási hiba

3.4. Objektumokat lépcsőzetesen a tanulási mintából kizáró modell-láncok

Lépések:

- OAM1 = 1254(O)·52(X)+1(Y), vagyis a teljes tanulási minta egyszerre került átadásra
- 9 pixelre sűrítve (1254 rangsorszám helyett)
- a COCO MCM elemző robotnak
- úgy, hogy a jó félvezetők Y értéke 1000, a rossz félvezetők Y-értéke 9000 volt
- majd a modell-output 9 pixeléből kizárásra került a csak jó félvezetőkre utal 1-2-3-4-es sáv
- így maradt 390 objektum (303+87) – mintha a jó réteg lehámozásra került volna (pl. egy almáról)
- a 390 objektum ismét átadásra került a COCO MCM számára
- a modell-outputok továbbra is 9 pixeléből (sávjából) kizárásra került a továbbra is csak jó félvezetőket tartalmazó 1-2-3-4-5-ös sáv
- a maradék 144 objektumban már csak 57 jó és továbbra is 87 rossz félvezető adata volt

- a 144 elemű tanulási halmaz lefutott 1000:9000-es (jó:rossz), ill. 99000:9000 (jó:rossz) és 1000:99000-es (jó:rossz) Y-értékpárokkal a sampling-et helyettesíteni hivatott Y-arányok és/vagy sorrendek hatásmértékének demonstrálása céljából, ahol
    - az 1000:9000-es klasszifikáló erejét leíró veszteség ráta (vagyis a tévesen pozitívnak vélt objektumok aránya egy helyesen pozitívnak vélt tesztobjektumra vonatkoztatva): 15.25 volt a kezdeti 17.41 helyett
    - a 99000:9000-es irányfordítás hatása masszív romlás volt: 22.71
    - míg az 1000:99000-es arány klasszifikáló erejét leíró veszteség ráta 12.28-ra csökkent…

Megjegyzések:

- a veszteségráta önmagában nem elég annak megítéléséhez, milyen jó egy modell a szakirodalmi megoldásokkal és a teoretikus benchmark-kal összevetve
- teoretikus benchmark-ként értelmezhető a legjobb = hibátlan megoldás is, melynek létét a saját 1567 objektumos tanulás is igazolja (ahol az első futtatás eredményeként 3/1567 hiba állt elő csak jó félvezetőket érintően az alább modell-struktúrával: 2 rétegben (1567 rekord / 52 attribútum / 5+1 modell / 26 pixel / COCO-MCM)
- a hibátlan modell feltételezése, mint best-of-benchmark felívja a figyelmet arra is, hogy az egymással teoretikus és best-of-benchmark nélkül versengő modellek, ha nem tudnak a kiválóság struktúrájáról semmit, akkor egymáshoz képest speciális versenyelőnyöket mutathatnak fel, melyek a hatékonyság (tanulási minta nagysága) alapján még tovább alakulnak – amit már a teoretikus és a best-of-benchmark nem képes irányítani…

4. EREDMÉNYEK

Logikai benchmark:

A logikai benchmark nem más, mint az, ha minden hibás félvezetőt hibátlannak tekintünk, hiszen olyan kicsi a hiba-arány, hogy pl. az accuracy így is 93% - vö. O12 TP, TF, FN, TN értékei:

	tp	fp	fn	tn	Y0	becslés	típus	átlag
O1	4	6	5	5	1000	1004.4	LR	1004
O2	9	9	3	3	1000	1000.4	ANN	1000
O3	9	9	6	6	1000	994.4	DT	996
O4	7	7	3	3	1000	1004.4	RF	1004
O5	2	3	8	8	1000	1003.4	LR	
O6	1	2	11	11	1000	999.4	ANN	
O7	7	7	6	6	1000	998.4	DT	
O8	9	9	2	2	1000	1003.4	RF	
O9	5	4	10	10	1000	995.4	coco(*)	997
O10	3	1	12	12	1000	999.4	coco	
O11	5	4	9	9	1000	997.4	coco	
O12	12	12	1	1	1000	999.4	teoretikus	999.4

11. ábra: Eredmények I. - sorszámozott nézet - forrás: saját ábrázolás

A best-of-objektummal kiegészített elemzések eredményei értelmében a nemzetközi szakirodalom (70:30 szcenárió) eredményei alig térnek el a normától. A két nézet kétféle lépcsős függvény triplet által került katalizálásra:

ABC-coco	tp	fp	fn	tn	Y0	Becslés	típus	átlag	
o1	5	7	5	5	1000	1003.7	LR	1003.0	
o2	10	10	3	3	1000	999.7	ANN	998.7	
o3	10	10	6	6	1000	993.6	DT	995.6	
o4	8	8	3	3	1000	1003.7	RF	1002.7	
o5	3	4	8	8	1000	1002.2	LR		
o6	2	3	11	11	1000	997.6	ANN		
o7	8	8	6	6	1000	997.6	DT		
o8	10	10	2	2	1000	1001.7	RF		
o9	6	5	10	10	1000	994.1	coco(*)	994.3	
o10	4	2	12	12	1000	994.6	coco		
o11	7	6	9	9	1000	994.1	coco		
o12	13	13	1	1	1000	995.6	teoretikus	995.6	
o13	1	1	1	1	1000	1021.8	best of	1021.8	<--genetikai potenciál

	tp	fp	fn	tn	Y0	Becslés	típus	átlag	
o1	5	7	6	6	1000	1003.3	LR	1002.6	
o2	10	10	4	4	1000	999.3	ANN	997.5	
o3	10	10	7	7	1000	993.2	DT	995.2	
o4	8	8	4	4	1000	1003.3	RF	1002.3	
o5	3	4	9	9	1000	1001.8	LR		
o6	2	3	12	12	1000	995.7	ANN		
o7	8	8	7	7	1000	997.2	DT		
o8	10	10	3	3	1000	1001.3	RF		
o9	6	5	11	11	1000	993.7	coco(*)	994.5	
o10	4	2	13	13	1000	994.2	coco		
o11	6	5	10	10	1000	995.7	coco		
o12	13	13	1	1	1000	996.2	teoretikus	996.2	
o13	1	1	1	1	1000	1025	best of	1025.0	<--genetikai potenciál

12. ábra: Eredmények II. - sorszámozott nézet - forrás: saját ábrázolás

A lépcsős függvények véletlen beállítású eredményei ebben a nézetben kevésbé előnyösek, mint az egymással versengő modellek nézetében, hiszen a nagy tévesen pozitívnak vélt objektum-arány a bets-of-hoz képest természetesen azonnal hátrányosnak tűnik:

Nyers adatok:

	tanulás	tp	fp	fn	tn	tp/fp	tp/fn	tn/fp	tn/fn	y0		fn/tp
o1	1000+	14.0	16.0	61.0	377.0	0.9	0.2	23.6	6.2	1000		4.4
o2	1000+	5.0	25.0	27.0	411.0	0.2	0.2	16.4	15.2	1000		5.4
o3	1000+	5.0	25.0	69.0	369.0	0.2	0.1	14.8	5.3	1000		13.8
o4	1000+	7.0	23.0	27.0	411.0	0.3	0.3	17.9	15.2	1000		3.9
o5	1000+	17.0	13.0	107.0	331.0	1.3	0.2	25.5	3.1	1000		6.3
o6	1000+	18.0	12.0	173.0	265.0	1.5	0.1	22.1	1.5	1000		9.6
o7	1000+	7.0	23.0	69.0	369.0	0.3	0.1	16.0	5.3	1000		9.9
o8	1000+	5.0	25.0	20.0	418.0	0.2	0.3	16.7	20.9	1000		4.0
C:	87+87	12.0	13.5	160.0	284.0	0.9	0.1	21.0	1.8	1000		13.3
A:	1000-	16.5	9.0	205.5	238.5	1.8	0.1	26.5	1.2	1000		12.5
B:	144	10.5	15.0	129.0	315.0	0.7	0.1	21.0	2.4	1000		12.3

13. ábra: Eredmények III. – Nyers adatok – forrás: saját ábrázolás

Rangsorszámok és irányok (0 = minél nagyobb, annál előnyösebb, 1 = minél kisebb annál előnyösebb)

irány	1	0	1	1	0	0	0	0	0	
	tanulás	tp	fp	fn	tn	tp/fp	tp/fn	tn/fp	tn/fn	y0
o1	4	4	6	4	4	5	3	3	4	1000
o2	4	9	9	2	2	9	4	9	2	1000
o3	4	9	9	5	5	9	11	11	5	1000
o4	4	7	7	2	2	7	1	7	2	1000
o5	4	2	3	7	7	3	5	2	7	1000
o6	4	1	2	10	10	2	6	4	10	1000
o7	4	7	7	5	5	7	7	10	5	1000
o8	4	9	9	1	1	9	2	8	1	1000
C:	2	5	4	9	9	4	10	5	9	1000
A:	3	3	1	11	11	1	9	1	11	1000
B:	1	6	5	8	8	6	8	6	8	1000

14. ábra: Eredmények IV. – Rangsorszámok – forrás: saját ábrázolás

Modellezett állapot (Excel-Solver-rel)

												hiba		
OAM	tanulás	tp	fp	fn	tn	tp/fp	tp/fn	tn/fp	tn/fn	y0		841		
return	2	3	4	5	6	7	8	9	10	y0	yb	delta	típus	átlag
o1	7	125	123	126	126	124	126	126	126	1000	1010	-10	LR	-10
o2	7	120	120	128	128	120	125	120	128	1000	997	3	ANN	1
o3	7	120	120	125	125	120	118	118	125	1000	979	21	DT	16
o4	7	122	122	128	128	122	128	122	128	1000	1008	-8	RF	-5
o5	7	127	129	123	123	126	124	127	123	1000	1010	-10	LR	
o6	7	128	130	120	120	127	123	125	120	1000	1001	-1	ANN	
o7	7	122	122	125	125	122	122	119	125	1000	990	10	DT	
o8	7	120	120	129	129	120	127	121	129	1000	1003	-3	RF	
C:	16	124	128	121	121	125	119	124	121	1000	1000	0	coco	-1
A:	8	126	131	119	119	130	120	130	119	1000	1004	-4	coco	
B:	17	123	126	122	122	123	121	123	122	1000	1000	0	coco	
total										11000	11000	0		

15. ábra: Eredmények V. – Modellezett állapot – forrás: saját ábrázolás

Az egymással versengő modellek (az anti-diszkriminatív elemzést nem tájékoztatva a teoretikus és a best-of állapotokról) azt jelzik, hogy a sok felesleges meggyanúsítás kompenzálható a nagyon leredukált input-adatmennyiség hatékonyságnövelő hatásával (vö. tanulás-oszlop (15. ábra:)).

5. SPECIÁLIS KIHÍVÁS – ANTI-DISZKRIMINATÍV MODELLEZÉS

Alternatív megoldások (objektum) összes elképzelhető jóságmutatója legyen Xi-attribútum. Alternatívák száma: minél több, annál jobb (vö. jövőkép, mint manuális alternatíva-gyár)

Y = Y0, azaz konstans idealitás-index (pl. 1000 jóságpont)

Feladat: MYX-FREE-COCO Y0 környezetben lehet-e minden objektum másként egyformán előnyös?

A speciális kihívás inputját a 80-20-as forrás-dokumentum 7. és 8. (6. ábra:és 7. ábra:) ábrája is adhatná megfelelő numerikus adatok esetén, ahol 3 dimenzió mentén az alábbi OAM áll rendelkezésre:

- Objektumok (3·3·4=36):
  - Minta alapján:
    - Unsampled
    - Oversampled
    - Undersampled
  - Klasszifikáció:
    - GBT
    - RF
    - LR
  - Szelekció:
    - Boruta
    - Mars
    - PCA
    - Worst case

- Attribútumok (1-2-3-4-5)
    - Accuracy
    - F-measure
    - Recall
    - Precision
    - FPR
    - (Y0)

Az előző fejezet teoretikus és best-of objektumokkal és azok nélkül értelmezett eredményei a 70:30-as objektum kontingencia-értékeihez képest értelmezte a lépcsős függvényeket csak találati arányok formájában és az ezek megtalálásához felhasznált objektumok számát minél kisebb annál jobb elvárásként leképezve.

6. KONKLÚZIÓK

A rendelkezésre álló adatvagyon kevesebb, mint 10 %-a került felhasználásra (vö. 591>52).

A hibátlan termékek tanulási outputja: tömegesen egyetlen egy érték!

A hibás és a hibátlan félvezetőket reprezentáló fiktív Y értékek aránya és sorrendje jelentősen hat a klasszifikációs potenciálra.

A modellek hibridizálása jelentősen hat a klasszifikálás sikerességére.

A pixelesség egyszerre erősíti és gyengítik a modellek robosztusságát, s mint ilyen hatásmechanizmus a jövőben optimalizálandó.

A legfrissebb eredmény esetén a korábban ismertetett módon megmaradt 144 elemből álló tanulási minta esetében pixelesség nélkül tanulás hibátlan. A teszt eredmény jelentősen javult a bemeneti jel sűrítése nélkül. A 80:20-as tanulás:teszt arány esetén 6 TP megtalált hibás félvezetőre már csak 61 FN jutott, ami jelentős javulást jelent a FN/TP mutatóban. Az alábbi ábra a modellek versenyét mutatja, amiben már 4. helyezést ért el a COCO az eddigi 7.hez képest.

	tp	fp	fn	tn	y0	becslés		sorrend	fn/tp
o1	14	16	61	377	1000	1004.7	LR	1	4.357143
o4	7	23	27	411	1000	1004.2	RF	2	3.857143
o5	17	13	107	331	1000	1004.2	LR	2	6.294118
o8	5	25	20	418	1000	1003.2	RF	4	4
o144_max_total_1	12	13.5	91	353	1000	1003.2	coco	4	7.583333
o2	5	25	27	411	1000	1000.2	ANN	6	5.4
o6	18	12	173	265	1000	999.2	ANN	7	9.611111
o10	16.5	9	205.5	238.5	1000	999.2	coco	7	12.45455
o7	7	23	69	369	1000	997.7	DT	9	9.857143
o11	10.5	15	129	315	1000	996.2	coco	10	12.28571
o9	10.5	15	141	303	1000	994.2	coco	11	13.42857
o3	5	25	69	369	1000	993.7	DT	12	13.8

16. ábra: Eredmények VI. –Legfrissebb – forrás: saját ábrázolás

7. JÖVŐKÉP: OPTIMALIZÁLÁS RÉTEGEI

Input-alternatívák:

A következő körben bevonható minden további attribútum, ahol pl. a 313 tesztesetre nincs adathiány, ha kihagyjuk a tanulási rekordok közül az adathiánnyal egyszer is érintett objektumokat – ha nem kívánunk adathiánypótlással semmilyen formában foglalkozni.

Optimalizálási alternatívák: [7]

8. FELHASZNÁLT FORRÁSOK

[1] SECOM-adatbázis: https://archive.ics.uci.edu/ml/datasets/SECOM

[2] Moldovan, Dorin & Cioara, Tudor & Anghel, Ionut & Salomie, Ioan. (2017). Machine learning for sensor-based manufacturing processes. DOI: 10.1109/ICCP.2017.8116997. https://www.researchgate.net/profile/Ionut-Anghel-3/publication/321260510_Machine_learning_for_sensor-based_manufacturing_processes/links/5ff5874392851c13feeff32b/Machine-learning-for-sensor-based-manufacturing-processes.pdf

[3] Jaekwon, Kim & Youngshin, Han & Jongsik Lee. (2016). Data Imbalance Problem solving for SMOTE Based Oversampling: Study on Fault Detection Prediction Model in Semiconductor Manufacturing Process. https://dokument.pub/data-imbalance-problem-solving-for-smote-based-flipbook-pdf.html

[4] Pitlik László, Pitlik Marcell (2021) Az attribútumok valódi értékének speciális összefüggései parallel termelési függvények generálásakor: https://miau.my-x.hu/miau/274/real_values_of_attributes.docx

[5] Pitlik László, Pitlik Marcell (2021) Roo-Boo-MOOC, avagy a robotizált MOOC-fejlesztés lehetőségei és korlátai II.: https://miau.my-x.hu/miau/272/Roo_Boo_Mooc_2.docx

[6] Számításokhoz használt online hasonlóságelemzési motor: https://miau.my-x.hu/myx-free/coco/

[7] Pitlik László (2021) MIAÚ ID_24481 Internet: https://miau.my-x.hu/miau/275/coco_mcm_opt_perx.xlsx

[8] Pitlik Marcell TDK 2020, Térfogati égés optikai vizsgálata és mesterséges intelligencia alapú elemzése https://combustioninstitute.hu/tdk_dij/2021/PitlikMarcell_TDK2020.pdf

[9] Andrew Ng (2021) MLOps: From Model-centric to Data-centric AI https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf https://www.youtube.com/watch?v=06-AZXmwHjo

9. MELLÉKLETEK

https://miau.my-x.hu/miau/275/felvezető*.xlsx

https://miau.my-x.hu/miau/275/

A COCO modellek mögötti egyenletek megtalálhatóak egy általam írt TDK [8] mellékletében egy példán keresztül részletesen bemutatva.