[ξ]

KSZI [ξ] AKTÁK

MTA KSZI műhelytanulmányok

MTA KSZI

Módszertani kísérlet az életpálya fogalmának formalizálására

– Előtanulmány a fiatal biológusok életpályakutatását célzó támogatott projekthez–

Soós Sándor

ssoos@colbud.hu;

2009/9

➲ http://www.mtakszi.hu/kszi_aktak/

Módszertani kísérlet az életpálya fogalmának formalizálására[1]

– Előtanulmány a fiatal biológusok életpályakutatását célzó támogatott projekthez–

Az életpályamodell formális megközelítése

Az életpályakutatás általános módszertanát illetően alapvető kérdés az életpálya fogalma, annak formalizálhatósága. Célunk az alább ismertetett munkában, amely a fiatal biológusok életpályakutatását célzó programhoz kapcsolódik, egy olyan modell vagy modellcsalád leírása volt, amely az életpálya-fogalmat formálisan és empirikusan is tesztelhető módon reprezentálja. Az ennek eredményeként előállt eszközöket arra az adatbázisra alkalmaztuk, amely a kutatás során fiatal biológusok körében végzett kérdőíves felmérés alapján keletkezett[2]. Ennek fő változói a karrierállomásokat, az életpályára hatást gyakorló számos további tényezőt (tanulmányok, családi körülmények stb.) és attitűdöket rögzítették.

A szokásos szóhasználattól kissé eltérően életpályamodell alatt olyan formalizmust, matematikai konstrukciót értünk, amely az alábbi jellemzőkkel rendelkezik: (1) modellezi, ill. reprezentálja az egyéni életpályák felvételéből, az adatbázisból kiemelkedő tendenciákat, (2) modellezi az életpálya alakulását befolyásoló tényezők kapcsolatát, és ennek révén (3) prediktív, vagyis előrejelzésre alkalmas.

A karrierkimenetek előrejelzési modellje: Bayes-klasszifikáció

A fenti három kritérium alapján az életpálya intuitív fogalmának megragadásához az ún. Bayes-hálók, ill. Bayes-klasszifikáció módszerét vizsgáltuk meg. A Bayes-háló általánosságban egy irányított (ciklusmentes) gráf, grafikus statisztikai modell, amely (többnyire) diszkrét valószínűségi változók együttes eloszlását reprezentálja. Csomópontjai az érintett változók, élei pedig a köztük fennálló feltételes függőségi viszonyokat jelzik: bármely gráfbeli változó feltételesen függ a szülőnódusoktól, vagyis azoktól a változóktól, és csak azoktól, amelyektől a gráfban él vezet ahhoz. Fontos kitétel a kizárólagosság: a vizsgált változók körében a szülőnódusoktól való függés a többi változótól való függetlenséget implikálja. Egy adott eredményváltozóra hatást gyakorló tényezők modellezésénél tehát egy Bayes-háló a (mért) háttérváltozók teljes összefüggésrendszerének tanulmányozását lehetővé teszi.

A Bayes-hálók egyik definitív jellemzője tehát a függőségi struktúra. A másik összetevő az egyes függőségi viszonyokat számszerűsítő paramétereket, vagyis a változóknak a szülőnódusoktól függő feltételes eloszlásait tartalmazza. A gyakorlatban egy X változó eloszlását ún. feltételes valószínűségi táblákkal (FVT) szokás meghatározni, amelyek minden kovariáns, vagyis a szülőnódusok lehetséges értékeinek

[1] A tanulmány eredményei a következő közleményben kerültek felhasználásra: Mosoniné Fried Judit — Pálinkó Éva — Soós Sándor: Tudományos fokozattal rendelkező fiatal biológusok munkahelyi orientációja. LVII. évf., 2010. január (71—90. o.) , Melléklet: Innovációkutatás.

[2] Az adatbázis jellemzését lásd az idézett cikkben.

minden kombinációja mellett rögzítik *X* kimenetét. A gráf és paraméterei ismeretében lehetségessé válik az eredményváltozó valószínűsíthető értékére vonatkozó következtetés, vagy, más aspektusból, az egyedek valószínűségi osztályozása, klasszifikációja.

Ez utóbbi elvet hasznosítják az alább bemutatott életpályamodellben is alkalmazott módszer, az ún. *Bayes-háló alapú klasszifikáció* során. A módszer feltételez egy adatbázist, amelynek egy választott *C* eredményváltozóját, az ún. osztályozó változót kívánjuk a többi változó (az ún. attribútumok) alkalmas Bayes-hálója segítségével modellezni. Az adatokhoz leginkább illeszkedő hálózat az ún. klasszifikátor *(classifier)*, amelynek fő funkciója, hogy lehetővé tegye az adatbázisban nem szereplő egyedek besorolását a *C* kategóriáinak valamelyikébe azok modellbeli attribútumai alapján. Ez a valószínűségi következtetés a Bayes-háló fent leírt tulajdonságain, és a klasszifikátorok általános felépítésén alapszik. Utóbbi alapvető vonása, hogy kiinduló- (gyökér-) pontjában a *C* áll: ebből azokhoz a változókhoz vezet él, amelyek közvetlenül függenek a *C*-től. Az indirekt hatású változókhoz más változóktól is futnak élek: ezek hatása tehát más változók értékétől is függ. E hatásokat az FVT-k számszerűsítik, amelyek becslése után a következtetés (bármely egyed *C* szerinti jellemzése) az ún. Bayes-szabályra támaszkodik, amely a megfigyelt X_1 … X_1 változók értékének függvényében úgy definiálja a *C* értékeinek feltételes, posterior valószínűségét, hogy az X_1 … X_1 értékeinek (kovariánsainak) a *C*-től való függésére – vagyis az FVT-kre – hivatkozik:

$$P(C = c_i \mid X_1 = x_1 \ldots X_n = x_n) = \frac{P(C = c_i)P(X_1 = x_1 \ldots X_n = x_n \mid C = c_i)}{P(X_1 = x_1 \ldots X_n = x_n)}$$

Ebből az elvből – mivel a Bayes-gráfalkotás szabálya szerint a kapcsolatban nem álló csomópontok függetlenek – levezethető, hogy a C-beli kategóriák valószínűsége az egyed $x_1 \ldots x_n$ tulajdonságai feltételes valószínűségének szorzata. Ezeket a valószínűségeket az FVT-k kódolják (direkt hatású változóknál a *C* értékeinek, indirekt hatásúaknál pedig a gráf szerint arra hatást gyakorló változók kovariánsainak függvényében).

A Bayes-klasszifikátor mint életpályamodell

Mindezek alapján a Bayes-háló alapú osztályozás ígéretes eszközt szolgáltat egy prediktív potenciállal rendelkező életpályamodell kidolgozásához, ha azt az alábbiak szerint közelítjük meg.

– Az karrier jelenlegi állapotát/kimenetét befolyásoló tényezőket és azok kapcsolatrendszerét kívánjuk hatékonyan modellezni.

– A feladatot klasszifikációs problémaként kezeljük: a kérdés, hogy milyen előrejelzést tehetünk a karrier kimenetét jellemző *C* változó (pl. beosztás, munkaerőpiaci pozíció, tudományos ranglétrán elfoglalt hely stb.) értékére a többi változó (előzmények, attitüdök stb.) értékeinek ismeretében.

Ebben a megközelítésben életpálya-modellünk egy $\langle G,\Theta \rangle$ Bayes-klasszifikátor, amelyben G egy irányított (körmentes) gráf C gyökérponttal (karrierkimenet), Θ pedig az ábrázolt eloszlás paramétereinek halmaza (FVT-k).

Egy ilyen életpálya-modell konkrét adatbázis alapján való feltárása az adatokhoz legjobban illeszkedő Bayes-klasszifikátor felépítését jelenti. Ennek korszerű módszere az adatbányászat és a mesterséges intelligencia területének metszetében álló *gépi tanulás* paradigmája. Az erre a célra készült algoritmusok az adatbázis alapján igyekeznek „megtanulni" a legjobb modellt. A Bayes-háló tanuló algoritmusok két komponense a megfelelő gráf, vagyis a változók függőségi struktúrájának keresése, illetve a struktúra alapján a paraméterek becslése. A tanulási folyamat vázlatosan egy kiinduló gráf heurisztikus ismételt módosítása valamely célfüggvény maximalizálása mellett, amely a gráfnak az adatokhoz (az együttes eloszláshoz) való illeszkedését méri *(network score)*. Az adatbázis egyidejűleg ún. tanító és tesztadatbázisként funkcionál: a tanuló algoritmus ez alapján konstruálja meg az osztályozó változó és az attribútumok közti viszonyokat, de – erre alkalmas érvényességi feltételek mellett – ezen is teszteli és értékeli a modellt (l. lent).

A következőkben bemutatott kísérlet célja egy ilyen életpályamodell, azaz (legalább) egy Bayes-háló alapú klasszifikátor (BHK) megalkotása a fiatal biológusokra vonatkozó adatbázis mint tanító- és tesztadatbázis felhasználásával. Az eredményként remélt predikciós eszköz döntéshozatali alkalmazásának előnyeit nem elsősorban a tényleges előrejelzésben – vagyis az új egyedek életpálya-kimenetének becslésében –, hanem a klasszifikátor tulajdonságainak jelentőségében jelölhetjük meg: predikciós modell „fehér doboz" jellegű, vagyis a predikció levezetése átlátható, és a hátterében álló struktúra betekintést nyújt a releváns életpálya-tényezők kapcsolatrendszerébe (egy adott kimeneti változó függvényében).

A fiatal biológusokra vonatkozó modell előkészítése: az adatbázis exploratív vizsgálata

Az adatbázisra épülő modellépítés alapvető mozzanata az életpálya-kimenet valamely releváns aspektusát kódoló C eredményváltozó kiválasztása. Nagyszámú potenciális magyarázó változó esetén – a Bayes-hálók számításigényét figyelembe véve – ugyancsak elkerülhetetlen a dimenzióredukció, vagyis a változóhalmaz jóval kisebb számú változóval való reprezentációja. Az általunk használt kérdőíves adatbázis – az aktuális kérdésfeltevéshez mért manuális szűkítés, vagyis a releváns dimenziókra vonatkozó minimalista hipotézis nyomán – nagyjából ötven kategoriális változóból épült fel. A C és a releváns attribútumok kiválasztásához egy exploratív technika vezetett el.

Az exploratív szakaszban az adatbázis változói közötti asszociációt és annak mértékét vizsgáltuk. Asszociációs mutatóként a kategoriális változók kölcsönös függőségét számszerűsítő ún. *kölcsönös információt* vagy MI-indexet használtuk, amely a változók együttjárását lényegében együttes eloszlásuk „egyenletességével", entrópiájával fejezi ki:

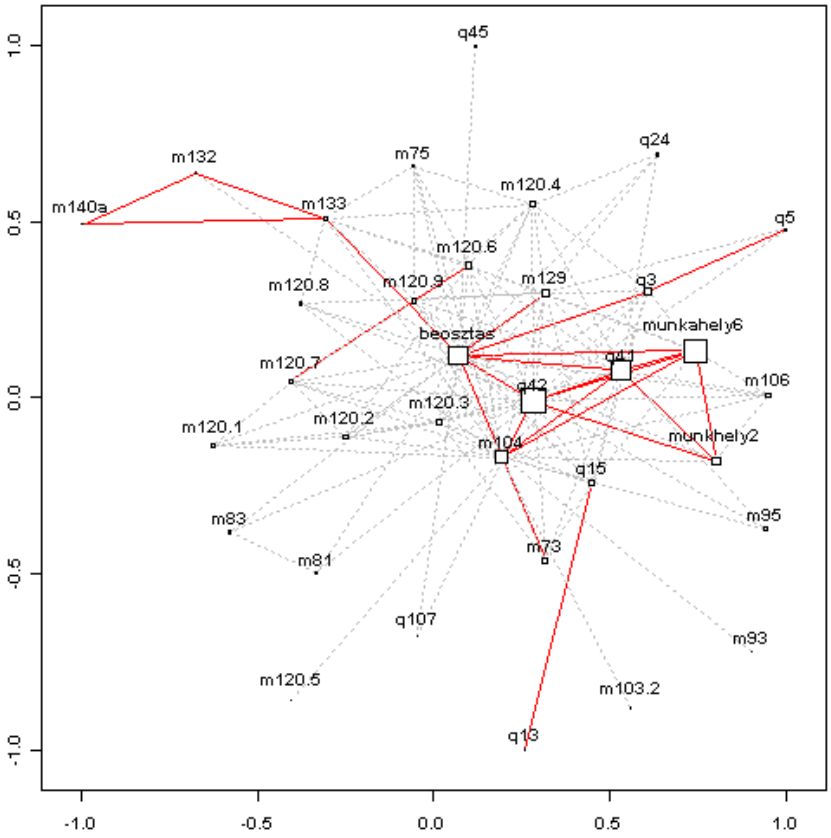
$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \text{ ahol}$$
$$H(A) = -\sum_i P(A_i) * \ln P(A_i)$$

az A változó entrópiája, P(A_i) pedig az A_i változóérték – együttes entrópia esetén az A_iB_i kovariáns – valószínűsége. Az ötven változó asszociációs mátrixa alapján ezek után egy asszociációs térképet készítettünk: egy olyan gráfot, amelynek csomópontjai a változók, élei pedig a köztük lévő asszociációs kapcsolatot jelzik. A gráfból töröltük az izolátumokat, vagyis azokat a változókat, amelyek kevésbé informatívak a többi viszonylatában. Az így megmaradt összefüggő részgráfot – komponenst – jeleníti meg a 1. ábra. Az asszociációs struktúra értelmezését két további jellemző segíti: 1) a mértékre vonatkozóan megállapítottunk egy határértéket (az asszociációra kapott hatványfüggvény-eloszlás alapján 0.2), és az ezt meghaladó kapcsolatokat folytonos vonallal, az alatta maradókat pontozott vonallal jelöltük. 2) A csomópontok nagysága azok ún. sajátvektor-centralitásával arányos: ez a változók pozíciójára vonatkozó mutató ebben az esetben úgy értelmezhető, mint az adott változó (közvetlen és közvetett) függőségi kapcsolatrendszerének mérete, kiterjedtsége.

Az gráf alapján megállapítható, hogy a fiatalbiológus-adatbázisban a változók közötti kapcsolat általánosságban viszonylag gyenge. Az általában látható gyenge kapcsolat mellett a változótérképről leolvashatók bizonyos, egymás tekintetében informatív változócsoportok. A térkép „magját” alkotó, viszonylag összefüggő csoport a munkapiaci státus különböző jellemzőit írja le: jelenlegi munkahely szektoriális helye (q42), szervezeti típusa (q41), beosztás (illetve két, ezekből képzett változó, amelyek definíció szerint összefüggenek az előbbiekkel: jelenlétük a módszer tesztjeként értékelhető). Hasonlóan erős klasztert alkot a családi paraméterek egy csoportja (családi állapot, háztartás mérete, van/nincs gyermek – m132, m133, 140a,). Figyelemre méltóbb ennek, valamint az előző klaszternek az összefüggése: a háztartás mérete (m133) és a beosztás között látható szorosabb összefüggés. További kapcsolatot figyelhetünk meg a fokozatszerzés átfogó területe (q5: infra- vagy szupraindividuális biológia) és annak intézménye között (q3, egyetemek): az utóbbi szintén kapcsolódik a beosztáshoz. A gráf tükrözte viszonylag triviális további összefüggés a képzés formája (q13) – nappali, levelező, esti tagozat – és a képzés finanszírozása között áll fenn (q15).

1.. ábra

A kérdőíves adatbázis változóinak asszociációs térképe (kölcsönös információ, MI alapján)



Figyelembe véve a változótérképet, Bayes-klasszifikátorunkhoz eredmény- vagy osztályozó változó gyanánt a jelenlegi munkáltató szervezeti típusát választottuk (q42), vagyis ugyanazt a dimenziót, amelyet korábban a karriertérképeknél is alkalmaztunk. Ez a tartalmi szempontból karrier-kimenet jellegű változó rendelkezik a legnagyobb (sajátérték-) centralitással, vagyis a legkiterjedtebb függőségi kapcsolatrendszerrel a változógráfban (a hasonló méretű „munkahely6” csomópont egy ebből képzett változót jelez).

A magyarázó változók kiválasztásához szintén egy MI-alapú modellépítési technikát, az adatbányászatban széles körben alkalmazott ún. korreláció-alapú jellemző-kiválasztást *(correlation-based feature selection, CFS)* hasznosítottuk. Az algoritmikus eljárás valamely keresési módszer segítségével igyekszik azt a változócsoportot azonosítani az adatbázisban, amely a legkevésbé redundáns és a leginkább informatív az eredményváltozó tekintetében. Ennek megvalósítása a heurisztikus keresés során megvizsgált változócsoportok értékelésével, és a legjobb jellemzőhalmaz kiválasztásával zajlik. Az alkalmazott kritérium az MI-index normalizált változata, az ún. *szimmetrikus bizonytalanság*, amely a kategoriális változók korrelációs mutatójaként szolgál: az ebből képzett maximalizált célfüggvény (merit) magas értéke azt jelzi, hogy a kiválasztott jellemzőhalmaz elemei a lehető legkevésbé korrelálnak

egymással (redundancia-mentesség), és a lehető legnagyobb mértékben a célváltozóval (informativitás).

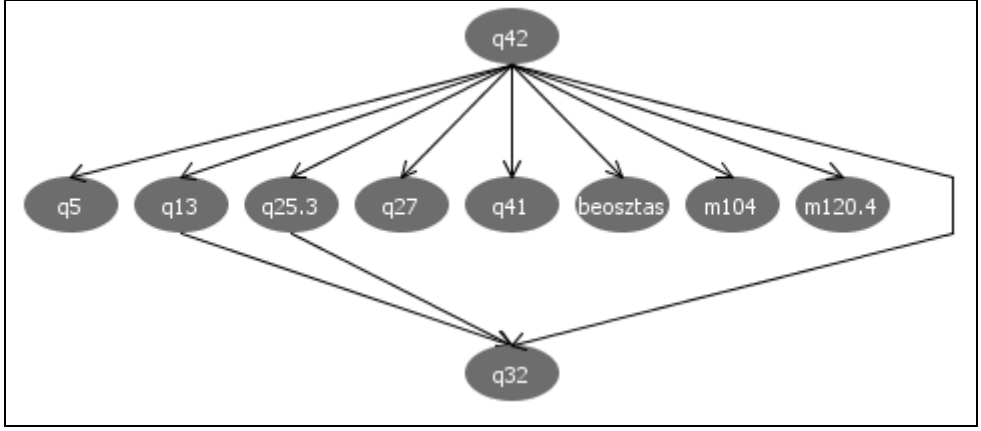
A klasszifikátor felépítéséhez a célváltozó (q42) megjelölésével a teljes adatbázison lefuttatunk egy CFS-keresést. Az így kapott legjobb változóhalmazra (és az eredményváltozóra) szűkített adatbázison gépi tanulással konstruáltuk meg a legvalószínűbb Bayes-hálózatot mint klasszifikátort. A feladathoz a Weka[3] nyílt forráskódú gépi tanulást megvalósító alkalmazáscsomagot vettük igénybe (a „Bayes net classifier" sémát a szülőnódusok lehetséges számát maximalizálva, egyébként az alapértelmezett paraméterekkel tanítottuk). Az alábbiakban elsőként röviden ismertetjük az ebből adódó modellt (2. ábra), majd annak értékelését.

A adatbázis alapján tanult modell

A különböző szervezeti típusokban való elhelyezkedés valószínűségét (q42) a mintából felépíthető modell értelmében kilenc változó befolyásolja. A változók tartalmát az 1. táblázatban foglaltuk össze. A hálózat struktúrája viszonylag egyszerű: egy tényező kivételével a csomópontok egyetlen szülője a célváltozó (q42), vagyis – ebben a kontextusban – egymástól függetlenül, közvetlenül hatnak a q42 kimenetére. A kivétel az a jellemző, amely a PhD-tanulmányok alatti külföldi szakmai tevékenységre vonatkozik (q32): ez (és így ennek hatása) két további tényezőtől függ: befolyásolja a képzés formája (q13: nappali, esti, levelező) és a doktori fokozat megszerzésére vonatkozó motiváció típusa (q25.3). Összességében azt mondhatjuk, hogy a kapott modellben a szervezeti hovatartozásra négy fő tényezőcsoporból következtethetünk: (1) A doktori képzésben való részvétel intenzitása: q13, q27, q32 (2) a munkapiaci előzmények: q41, (3) a tudományos karrierre vonatkozó attitűdök. q25.3, m104, m120.4, ill – sajátos, de érthető módon – a fokozatszerzés átfogó szakterülete.

[3] WITTEN, I. H.– EIBE , F. [2005]: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.

2. ábra
A munkáltató szervezeti típusát előrejelző Bayes-klasszifikátor



1. táblázat
A Bayes-klasszifikátort felépítő változók definíciója

Kód	**Leírás**
q5	Melyik átfogó területen szerzett fokozatot?
q13	Milyen tagozaton járt doktorképzésre?
q25.3	Az alább felsoroltak közül mi ösztönözte Önt leginkább a fokozat megszerzésére?
q27	Tanult, esetleg dolgozott Ön ösztöndíjasként külföldön egyetemista korában?
q32	Tanult, esetleg dolgozott Ön ösztöndíjasként külföldön PhD hallgató, illetve doktorjelölt korában?
q41	Milyen ágazatban tevékenykedik jelenlegi/idöben legközelebbi munkáltatója?
beosztas	Mi a beosztása?
m104	Életpályája során hogyan szeretné maximálisan kiteljesíteni saját tudományos-szakmai ambícióit? Kérjük, az alább felsoroltak közül válassza ki azt az állítást, melyet a leginkább sajátjának érez!
m120.4	Kit tart Ön sikeres kutatónak? / Magas presztízsü állást tölt be

Az összefüggésrendszer részleteit, mint korábban részleteztük, az FVT-k tartalmazzák, vagyis azok a feltételes valószínűségi táblák, amelyekből megállapítható, hogy a jellemzők egyes értékei milyen mértékben valószínűsítik az egyes szektorokat (mivel itt a legtöbb változónak egyetlen szülőnódusa van, a célváltozó, ezek a táblák egyszerű kereszttáblák: kivétel ez alól a három jellemzőtől függő q32, amelynek FVT-je az összes kovariáns, vagyis a három jellemző értékének összes kombinációja mellett adja meg ezeket a valószínűségeket). Ezek az FVT-k a modell részét képezik, a gráf egyes csomópontjaihoz tartoznak. Demonstrációs céllal itt a q5-höz tartozó FVT-t közöljük, amely a fokozatszerzés átfogó területének szerepét jellemzi (2. táblázat).

2. táblázat

A fokozatszerzés átfogó területéhez (q5) tartozó feltételes valószínűségi tábla (FVT)

	Infraindividuális biológia	**Szupraindividuális biológia**	**Nem dönthető el igazán**	**Más szakterületen, éspedig:**
Akadémiai kutatóintézet	0.57	0.19	0.12	0.12
Egyéni vállalkozás	0.25	0.25	0.25	0.25
Egyetem	0.53	0.35	0.08	0.04
Más állami K+F intézmény	0.63	0.25	0.00	0.13
Más intézmény, éspedig:	0.38	0.63	0.00	0.00
Nonprofit szervezet	0.00	0.67	0.33	0.00
Profitorientált kisvállalkozás (10-49 fös)	0.00	0.00	1.00	0.00
Profitorientált középvállalkozás (50-250 fös)	0.67	0.33	0.00	0.00
Profitorientált mikrovállalkozás (max. 9 fös)	0.00	1.00	0.00	0.00
Profitorientált nagyvállalat (250 fö felett)	0.86	0.14	0.00	0.00

A táblázatból tanúsága szerint ha egy biológusjelölt az infraindividuális biológiát választja, úgy jóval nagyobb a valószínűsége a profitorientált nagyvállalatnál való elhelyezkedésnek, mint a ha a szupraindividuális területen kutat. Ez az eredmény jól interpretálható, minthogy intuitíve is könnyebb biokémikusként vagy molekuláris biológusként ipari területen érvényesülni (gyógyszeripar), mint taxonómusként vagy ökológusként. A táblázat szerint ugyanakkor az akadémiai és az egyetemi szféra is „favorizálja” ezt a területet (bár az egyetemek esetében kiegyenlítettebb az arány). Másfelől a szupraindividuális terület elsősorban a nonprofit szférát (ill. a mikrovállalkozást) valószínűsíti: ez szintén értelmezhető, ha figyelembe vesszük a jellemzően ökológusokat igénylő környezetvédelem mint szektor szervezeti formáit. A konkrét magyarázatokon túl ugyanakkor figyelemre méltó, hogy az elhelyezkedési mintázatokat már a (bilógián belüli) területválasztás nagyban befolyásolja.

A modell értékelése

Az fent bemutatott modell érvényének vizsgálatához több, az adatbányászatban klasszifikátorok értékelésére jellemző mutatót vizsgáltunk meg. A modellnek az adatokhoz való lehető legjobb illeszkedését már maga a tanulási folyamat is igyekszik biztosítani: a gépi tanulás két lépcsője, a tanulás és tesztelés az ún. keresztvalidáció módszerével zajlik. (*Witten–Eibe*, [2005]). Ebből fakadóan minden egyed szerepel legalább egyszer tanító- és tesztalanyként is.

A modellépítés első lépcsőjeként értékelhető jellemző-kiválasztás (CFS) sikerességének mutatója igen magas (merit = 0.9). Némileg árnyalja ugyanakkor ezt a képet, ha a magyarázó változók halmazát összevetjük a változótérképpel. A magyarázó változók között megtaláljuk a térkép szerint a célváltozóval erősen asszociált (vele függőségi viszonyban lévő) változókat, de olyanokat is, amelyek a térkép szerint gyengébb kapcsolatban állnak vele. Másfelől az erősen asszociált magyarázó változók egy kisebb köre egymással is szorosabban korrelál (ami ellentmondani látszik a rendundancia-elvnek). Az első megfigyelésre a magyarázat az, hogy a térkép csak a legerősebb kapcsolatokat vizualizálja (amelyek sokszor viszonylag triviálisak). A második megfigyelés a függőségi viszonyrendszer egészével hozható kapcsolatba: a függőség mértékének ilyen eloszlása mellett (kevés erős, sok gyengébb kapcsolat) a legjobb változóhalmaz kiválasztása kompromisszumot eredményez. Összességében azt mondhatjuk, hogy a CFS nagyobb felbontású képet nyújt a releváns változókról.

A kísérlet eredményeként kapott klasszifikátor értékelésének elsődleges mutatója a megfelelően besorolt egyedek száma, illetve részaránya (*percent of correctly classified instances*, PCC). Az adott tanító (és teszt-) adatbázison a fenti, vagyis a legjobban illeszkedő modell közepes teljesítményt mutat: a munkáltató szervezeti típusát 63%-ban volt képes helyesen (vagyis az adatbázisban szereplő tényleges érték szerint) megítélni. Ez a részarány azonban egyidejűleg a célváltozó egyes értékeire, vagyis az egyes szervezeti típusokra vonatkozó helyesség, a helyes pozitív ráták átlaga. A klasszifikátor teljesítményét az egyes kimenetekre vonatkozóan értékelve jóval informatívabb képet kapunk (3. táblázat). A táblázat értelmében a modell kellően megbizható eredményt, vagyis jó jellemzést ad az *akadémiai kutatóintézet*, az *egyetem* és a *profitorientált nagyvállalat* típusra vonatkozóan, a többi esetében viszont kimondottan rosszul teljesít. Ez utóbbi értékek olyan, bővebb minta révén javíthatók, amelyben a vonatkozó kategóriákat jóval több eset képviseli, ebből következőleg a tanuló algoritmus számára több információ áll rendelkezésre az adott kimenetekről.

3. táblázat

A Bayes-klasszifikátor értékelése

Érték	**HP ráta**	**FP ráta**	**Pontosság**	**Fedés**
Akadémiai kutatóintézet	0.81	0.288	0.596	0.81
Egyetem	0.714	0.123	0.795	0.714
Más állami K+F intézmény	0	0.044	0	0
Egyéni vállalkozás	0	0	0	0
Profitorientált mikrovállalkozás (max. 9 fös)	0	0	0	0
Profitorientált kisvállalkozás (10-49 fös)	0	0	0	0
Profitorientált középvállalkozás (50-250 fös)	0	0.008	0	0
Profitorientált nagyvállalat (250 fö felett)	0.857	0.035	0.6	0.857
Nonprofit szervezet	0	0	0	0
Más intézmény, éspedig:	0.375	0.018	0.6	0.375
Súlyozott átlag (- PCC)	0.639	0.155	0.599	0.639

Rövidítések: HP=helyes pozitív, FP=false pozitív