Acta Agraria Kaposváriensis (2010) Vol 14 No 3, 281-291
Kaposvári Egyetem, Állattudományi Kar, Kaposvár
Kaposvár University, Faculty of Animal Science, Kaposvár

Bükkös erdő mikroklíma adatainak vizsgálata adatbányászati eszközökkel

Pödör Z.
Nyugat-magyarországi Egyetem, EMK, Matematikai Intézet, 9400 Sopron, Ady Endre út 5.

ÖSSZEFOGLALÁS

Az adatbányászat hatékonyan alkalmazható mindenütt, így az erdészetben is, ahol nagy adathalmazok állnak rendelkezésre. Egy bükkös erdő mikroklimatikai adathalmazát felhasználva elemezzük a sugárzási adatokat, vizsgáljuk, hogy milyen összefüggések állnak fenn ezek és a többi paraméter között - amelyeket hagyományos eszközökkel nem vagy csak nehezen lehet megtalálni - és ebből milyen következtetések vonhatóak le. Megmutatjuk, hogy az adatbányászat eszközei adathibák javítására, műszer meghibásodásból eredő adathibák felderítésére, javítására is alkalmazhatóak. Bemutatunk egy lehetőséget, melynek segítségével számszerűen is mérhetjük diszkretizált folytonos adatok esetén a jóslás pontosságát.
(Kulcsszavak: adatbányászat, erdő sugárzási paraméterei, hiba felderítés, adatjavítás)

ABSTRACT

Research the microclimate data of a beech wood with datamining toolbar
Z. Pödör
University of West Hungary, Faculty of Forestry, Mathematics Institute, H-9400 Sopron, Ady Endre út 5.

The datamining toolbar can be applied in areas, like forestry, where we have huge databases. We use a database of a beech wood, and we analyse the radiation parameters. We can easily lay down a map of the connection between these parameters and the other attributes. We can discover such relationship that cannot be found easily with traditional methods, or can't be found at all. We show the methods of datamining that are suitable for handling missing and false data. the hintten data errors can be also found, which generate the error of indicator. We demonstrate a method, with it support we can determine the goodness of prediction when we use continuous data.
(Keywords: datamining, radiation parameters of a wood, handling missing and false data, explorating dara errors)

BEVEZETÉS

Ma már szinte az összes tudományterület jellemzője, hogy hatalmas méretű adathalmazokkal rendelkezik egy-egy kutatás kapcsán. Azonban ezek így csak puszta adattemetők, ezeket fel kell dolgozni, ki kell nyerni a hasznos információkat. Erre rendelkezésünkre állnak a hagyományos statisztikai, matematika eszközök. A mai kutatások egy jelentős részére (pl. az erdészetben is) jellemző, hogy az adatokból a már jól bevált, elfogadott gondolatmenetek alapján állítunk fel összefüggéseket. Azaz sokszor a vizsgálatok megkezdésekor már tudjuk, sejtjük, hogy milyen típusú kapcsolatokat akarunk keresni. Erre a fent említett hagyományos apparátus tökéletesen

alkalmas. Azonban a jelentős méretű adatbázisok ennél sokkal több hasznos összefüggést, információt is tartalmazhatnak, amiknek felfedésére az adatbányászat lehet alkalmas (*Abonyi*, 2006; *Bodon*, 2008; *Han, Camber*, 2006). A vizsgált adatbázis egy bükkösben gyűjtött mérési adatokból áll, melyben adatbányászati eszközökkel elemezzük a mért a sugárzási adatokkal kapcsolatos összefüggéseket.

Ahhoz, hogy a kinyert információk, összefüggések helyesek, használhatóak legyenek feltétlenül szükséges, hogy a vizsgált adatbázis lehetőleg teljes, azaz hiány és hibamentes legyen. Ez a gyakorlatban általában nincs így, ezért megmutatjuk, hogy ha rendelkezésre áll egy megfelelő tanulóhalmaz, akkor az adatbányászat eszközei alkalmazhatóak hiányzó, hibás attribútum értékek pótlására, javítására. Bemutatjuk továbbá, hogy a műszer meghibásodásból adódó rejtett mérési hibák is felderíthetőek és javíthatóak az adathalmazban.

ANYAG ÉS MÓDSZER

Az adatbázis

A használt adatbázis egy 55-57 éves, középkorúnak tekinthető bükkösben végzett mérések adatait tartalmazza. A bükkös koronaszintje 15-19 méteres, zárt, alatta a törzstér üres. Az itt jelenleg is folyó kutatás alapvető célja az elsősorban a koronaszint által irányított légkörfizikai folyamatok vizsgálata. Ez egy 30 méter magas toronyba szerelt mérőműszerek segítségével történik. Ezek az eszközök 10 percenként több paramétert is mérnek különböző szinteken. A telepített szenzorok 30, 23, 19, 14 és 2 méteren vannak, ahol szintenként mérünk: hőmérsékletet (min., max., átlag, aktuális érték a mérés pillanatában); szél adatokat (átlag, max. szélerősség és irány); légnedvesség adatokat; fotoszintetikusan aktív sugárzást, valamint mérünk még (de már nem szintenként) légnyomás értékeket; 30 és 23 méteren sugárzási adatokat; talajhőmérsékletet 0, 5, 10, 20, 50, 100 cm mélyen; levélfelszín hőmérséklet számításához szükséges adatokat; hómagasságot; csapadékmennyiséget; a talajnedvesség számításához 10, 20, 30, 40, 60 és 100 cm mélyen paramétereket.

A mérések 2006.05.01-től 2008.08.01-ig álltak teljességgel a rendelkezésre 10 perces gyakorisággal. Komoly műszer meghibásodás nem történt, a néhány alkalommal fellépő adathiányosságokat könnyen tudtuk pótolni a szokásos statisztikai módszerekkel. Illetve néhány esetben - nem a műszerek hibájából - hiányoznak rekordok, de ezek mennyisége nem befolyásolja a vizsgálatot. Az adatgyűjtés során folyamatosan ellenőriztük az adatok helyességét (nemcsak egyszerűen a műszer által adott „valid" jelzést fogadtuk el). A közvetlenül mért paramétereken kívül egyéb, az alapkutatás szempontjából fontos számított adatokat is tartalmaz az adatbázis, mint pl. levél és virtuális hőmérséklet, albedo, páranyomás, stb.. Így összességében elmondhatjuk, hogy egy hibáktól mentes, teljesnek mondható adathalmaz áll rendelkezésre a fenti időszakból kb. 60 attribútummal és nagyságrendileg 115000 db rekorddal. Mindez megfelelő alapot jelent a bányászati modellek megalkotására és a korábban említett elemzések elvégzésére. A 2008.08.01-et követő időszakról is rendelkezünk már adatokkal, melyeken bemutatjuk az adatjavítás, mérési hibák felfedésének módszerét és eredményeit. A vizsgálatokhoz az MSSQL server 2008 szoftvert használjuk.

EREDMÉNY ÉS ÉRTÉKELÉS

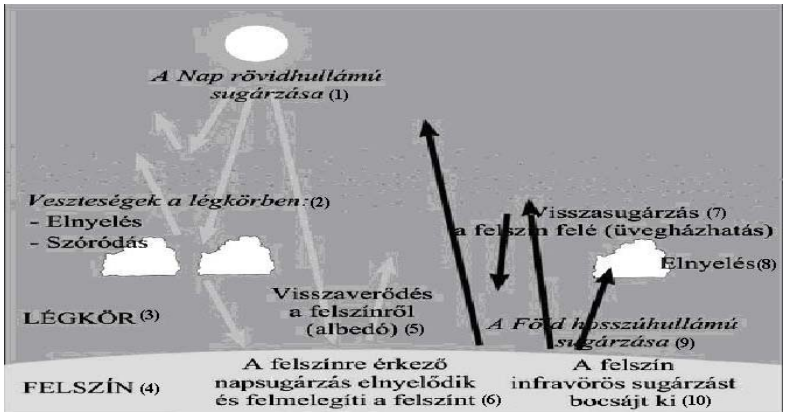
Adatok elemzése

Az adatok közötti kapcsolatok kezdeti, általános feltérképezésére egy megfelelő kiindulópont lehet a Naive Bayes algoritmus futtatása minden attribútumra, mint bemenő

és jóslandó paraméterre. Így adódik egy olyan gráf struktúra, amely adott erősségi szinten megmutatja az attribútumok közötti kapcsolatokat. Pontosabban azt, hogy az adott attribútum mely attribútumokat jósolja, illetve az adott attribútumot mely attribútumok jósolják az adott szinten. Természetesen lehetőségünk van ennél konkrétabb összefüggés vizsgálatra, ahol kiemelünk paramétereket és megvizsgáljuk, hogy ezek milyen kapcsolatban állnak a többivel. Ezt alkalmazva vizsgáljuk meg a sugárzási jellemzőket. Ehhez először röviden tekintsük át (*1. ábra*) a mért sugárzási paraméterek hátterét.

1. ábra

A sugárzás útja a légkörben



Forrás (*Source*): *Vig* (1995)

Figure 1. The way of radiation in the atmosphere

The short-wawe radiation of the Sun(1), Atmosphere demage(2), Atmosphere(3), Ground(4), Albedo(5), The incoming radiation is absorbed, and hots up the ground(6), Greenhouse effect(7), Absorbtion(8), The long-wawe radiation of the ground(9),The ground deflates infrared radiation(10)

A Nap irányából rövidhullámú (globális) sugárzás érkezik a felszín felé, mely a légkörben történő elnyelés és szóródás után éri el azt. Ennek a sugárzásnak egy része változatlan formában visszaverődik a felszínről és elhagyja a légkört. A beérkező és a visszaverődő rövidhullámú sugárzás különbsége a rövidhullámú sugárzási egyenleg, arányuk az albedo. A Föld felszín az elnyelt rövidhullámú sugárzás hatására felmelegedik és így már hosszúhullámú sugárzást bocsát ki. Ennek egy jelentős része visszaverődik a légkörből pl. az üvegházhatású gázok miatt, így felmelegítve azt. Ez egy szükséges jelenség, hiszen e-nélkül kb. 33 °C-kal lenne alacsonyabb a Föld átlagos hőmérséklete; azonban olyan méreteket öltött ez a folyamat, mely már veszélyezteti a globális klímát: felmelegedés, üvegházhatás. A felszín által kisugárzott és a légkör által visszavert hosszúhullámú sugárzás különbsége a hosszúhullámú egyenleg, továbbá a rövid és hosszúhullámú sugárzási egyenlegek különbsége a sugárzási egyenleg, amely meghatározza az éghajlati folyamatok energia forrását.

A sugárzási paraméterek vizsgálata

Az adatbázis a mérések alapján tartalmazza a be és kimenő hosszú és rövidhullámú sugárzás mértékét, valamint számított adatként a hosszú és rövidhullámú sugárzási egyenleget és az albedot. A vizsgálatok elvégzésére alkalmas eszközök az adatbányászat osztályozó módszerei: Naive Bayes, Döntési fa. Elemzéseinket a teljes, a lombos és lombtalan állapotra végezzük el, összevetve a kapott eredményeket és kiemelve a figyelmet érdemlőeket.

Természetesnek tekinthető, hogy a sugárzási paraméterek között szoros kapcsolat áll fent, így ezeket külön nem emeljük ki a vizsgálatban. A teljes időszak tekintetében a következőket mondhatjuk el: a bejövő globális sugárzással legszorosabb kapcsolatban álló paraméterek a levélfelszín hőmérséklet, valamint a 19 méter magasságban mért paraméterek, mint pl. hőmérséklet, légnedvesség. Ennek oka nyilván az, hogy a beérkező rövidhullámú sugárzás nagy részét a kb. 19 méteres magasságban elhelyezkedő lombkorona nyeli el (nyári időszakban ez 90-95% és még télen is 60% körül van). Megmutatható, hogy a légnedvesség értékek és a bejövő globális sugárzás között negatív korreláció áll fent (r=-0,45), ami arra utal, hogy ha esik az eső, azaz felhős az ég, akkor kisebb a bejövő globális sugárzás mértéke. Azonban a lombtalan időszakban a 19 méteren mért értékek már jóval gyengébb kapcsolatot mutatnak a bejövő globális sugárzással, hiszen a levélzet hiánya miatt a sugárzás az alacsonyabb szinteket is eléri, és ott hasznosítódik. A visszavert rövidhullámú sugárzás tekintetében hasonló kapcsolatokat tapasztalhatunk a többi paraméterrel. A bejövő globális sugárzás egy része egyszerűen visszaverődik a felszínről, és el is hagyja a légkört. Ezen visszaverődési arány mérőszáma az albedo, melynek mértéke a bükkös faj esetén 14-16% körül van.

Lombos időszak vonatkozásában a globális bejövő és kimenő sugárzás, valamint az ezekből számított albedo szoros kapcsolatban áll a 19 méteren mért szélerősséggel is. Ugyanakkor nem tapasztaljuk ezt a kapcsolatot a hosszúhullámú sugárzási paraméterekkel ebben az időszakban és egyetlen sugárzási paraméterrel sem mutat erősnek mondható kapcsolatot a lombtalan időszakban. Az összefüggések pedig azt mutatják, hogy erősebb szélben (*1. táblázat*) a bejövő és visszavert globális sugárzás mértéke nagyobb, ugyanakkor az albedo mértéke jelentősen csökken. Az albedo csökkenő mértéke valószínűleg avval magyarázható, hogy erősebb szélben a mozgó levelek miatt jobban áthatol a sugárzás a lombkoronaszinten és így arányaiban kevesebb sugárzás verődik közvetlenül vissza a légkör felé.

1. táblázat

A 19 méteren mért szélerősség és a globális sugárzás, valamint az albedo erősségének kapcsolata lombos állapotban

Szélerősség (1)	szel19<0,5	szél19<1	szel19<2	szél19>2
Bejövő glob. sugárzás (2)	223,9606	316,7522528	363,0909	522,5479
Kimenő glob. sugárzás (3)	32,81898	45,32922	51,6724	70,86815
Albedo (%) (4)	18,17805	16,8852	16,29924	13,47645

szel19: a 19 méteres magasságon mért szélerősség (m/s) (*The windforce (m/s) in 19 m*)

Table 1. The connection between the strongness of windforce, global radiation and albedo in 19 m. in the leaved status

Windforce(1), Incoming global radiation(2), Outgoing global radiation(3), Albedo %(4)

Megfigyelhetjük még, hogy a lombos erdő albedoja nagyobb mint a lombtalané (*2. táblázat*), a biztosan lombos időszakban az albedo átlaga 16,05%, míg a biztosan lombtalan időszakban 11,32%. Ugyanakkor a lombos időszakon belül vizsgálva az albedo változását az idő függvényében, adódik, hogy az őszi, de még lombos időszakban csökken a mértéke. Ez arra utal, hogy a lombkorona színváltása csökkenti a visszavert rövidhullámú sugárzás arányát.

2. táblázat

Albedo átlagos mértéke nyáron és őszi, lombos időszakban

Hónap (1)	**2006.07.**	**2006.09.**	**2007.07.**	**2007.10.**
albedo(%) átlag (2)	16,61544	15,20185	14,3723	13,47568

Table 2. The average measure of albedo in summer and autumn of leaved season

Month(1), Albedo(%) average(2)

A felszínről vissza nem vert globális sugárzás hasznosul valamilyen formában: transzspiráció (párologtatás), felszín, levegő felmelegedése. A felszín felmelegítése során az elnyelt rövidhullámú sugárzás átalakul és hosszúhullámú sugárzás formájában hagyja el a felszínt. Ez az, ami pl. az üvegházhatású gázoknak köszönhetően visszaverődik felmelegítve a légkört.

A kimenő és visszaverődő hosszúhullámú sugárzást vizsgálva (*3. táblázat*) azt tapasztalhatjuk a teljes időszak vonatkozásában, hogy a legszorosabb kapcsolatot mutató paraméterek a léghőmérséklet adatok, valamint a telítettségi páranyomás (tpny: az egységnyi térfogatú légoszlopban a vízgőz parciális nyomása) és páranyomás (pny: a levegőben levő vízgőz nyomása) értékek. Ennek oka, hogy, hogy a bejövő rövidhullámú sugárzás párologtatásra, illetve a felszín felmelegítésére fordítódik. Közelebbről vizsgálva a kapcsolatot látható, hogy ezek a légnedvességet jelző paraméterek erős pozitív korrelációban állnak a kimenő és erősnek mondható kapcsolatban a visszavert hosszúhullámú sugárzási paraméterekkel. A talajszinten mért hőmérséklet adatok is erősnek tekinthető (bár az előzőeknél gyengébb) kapcsolatban állnak a hosszúhullámú sugárzási paraméterekkel, hiszen ezek felelnek a felszín felmelegedéséért. Ugyanakkor éppen ez a gyengébb kapcsolat magyarázza azt, hogy nyáron az erdők belseje hűvös marad, hiszen a lomkorona jelentős sugárzáselnyelő képessége miatt kevesebb energia jut a talaj és így az erdő belsejének felmelegítésére.

A leírtak alapján arra következtethetünk, hogyha magas a levegő nedvességtartalma, akkor a növények párologtatási képessége csökken, így a bejövő rövidhullámú sugárzás kisebb része fordítódik transzspirációra és nagyobb része a felszín felmelegítésére. Ezért magasabb lesz a hosszúhullámú sugárzás mértéke is. Másképpen fogalmazva, ha intenzív a növényzet transzspirációja, akkor kevesebb energia jut a felszín melegítésére, így valamelyest csökken a hosszúhullámú sugárzás mértéke. A számított paraméterek vizsgálatára nem térünk ki, hiszen nem meglepő módon az alap attribútumoknál tapasztalt összefüggések állnak elő.

A fent leírtak alapján is látható, hogy egy ilyen típusú, méretű adatbázisban az adatbányászati eszköztár alkalmazásával felfedhetőek olyan összefüggések, kapcsolatok, melyeket a hagyományosnak tekintett eszközökkel nem, vagy csak nehezen lehetne előállítani. A felállított összefüggésekről eldöntjük, hogy lényegesek-e vagy sem, illetve

megpróbálunk magyarázatot találni a fontosnak vélt relációk okára. Ehhez elengedhetetlen az adott terület szakemberével a folyamatos együttműködés.

3. táblázat

Hosszúhullámú sugárzás korrelációs kapcsolata egyéb paraméterekkel

hosszhull. kisugárzás (1)	**hom19**	**hom23**	**tpny19**	**tpny23**	**talajhom5**	**levfelszho**
korreláció mértéke (r) (2)	0,993	0,990	0,973	0,971	0,912	0,997
hosszhull. visszavert sug. (3)	**pny14**	**pny2**	**pny23**	**talajhom0**	**talajhom5**	**hom30**
korreláció mértéke (r) (2)	0,756	0,758	0,758	0,688	0,689	0,644

hom19, hom23: 19 és 23 méteren mért léghőmérséklet (C°) (*airtemperature(C°) in 19 and 23 m*); tpny19 és tpny23: telítettségi páranyomás (Hgmm) 19 és 23 méteren (*saturation vapour pressure (Hgmm) in 19 and 23 m*); talajhom0, talajhom5: talajhőmérséklet (C°) 0 és 5 centiméter mélyen (*ground temperature (C°) in 0 and 5 cm deep*); pny2, pny14, pny23: páranyomás 2, 14 és 23 m magasan (*vapour pressure (Hgmm) in 2, 14 and 23 m*); hom30: léghőmérséklet 30 m magasan (*airtemperature in 30 m*)

Table 3. The correlation connection between long wave radiation and other parameteres

Long-wawe outgoing radiation(1), The measure of correlation coefficients(2), Long-wawe reflected radioation(3)

Adathibák

Az adatbányászat egyik alapfeladata, hogy új tudást, összefüggéseket nyerjünk ki az adatainkból. Ahhoz, hogy az összefüggések, információk valóban helytállóak lehessenek fontos, hogy az adatbázis lehetőleg minél teljesebb (ne legyenek benne hiányok, vagy minél kevesebb) és hibamentesebb legyen. Ennek biztosítása alapvetően az adatelőkészítés feladata, ami rendkívül sok energiát és időt is igénybe vehet a használni kívánt adatbázis állapotától függően. Azonban ha már rendelkezésünkre áll egy ilyen adathalmaz, akkor a későbbiekben az alap adatbázishoz kapcsolt újabb rekordok esetén alkalmazhatjuk az adatbányászat eszközeit arra, hogy hiányzó attribútum értékeket pótoljunk, hibás adatokat javítsuk, illetve felderítsük a műszer meghibásodásából eredő rejtett mérési hibákat és javítsuk azokat. Fontos megemlíteni, hogy nem teljes rekordok pótlásával foglalkozunk, hanem olyan esetekkel, amikor rekordokból néhány hiányzó, hibás attribútum értéket javítunk, pótlunk azonban akár hosszabb időintervallumon keresztül is. Ennek a lehetőségét, illetve az evvel kapcsolatos problémákat és lehetséges megoldásokat mutatjuk be a bükkös adatbázis kapcsán.

Mérési adatok lehetséges hibái

A természetben lejátszódó folyamatokat vizsgáló tudományterületek (pl. erdészet) jellemzője, hogy az adatbázisainak egy jelentős része mérőműszerek által mért mérések eredményeit tartalmazza. Ez alapvetően magában hordozza a hiba lehetőségét a műszer meghibásodások miatt. A szenzorok általában rendelkeznek egy ellenőrző rendszerrel, amely a mért adatot „valid", illetve „invalid" jelzéssel látja el attól függően, hogy helyesnek, vagy helytelennek tekintette. Vegyük pl. a levegőhőmérsékletet mérő szenzorokat: ezek elfogadási tartománya magyarországi viszonylatban pl. [-25,45] intervallum lehet. Így ha egy nyári nap folyamán az eszköz -10 °C-ot mér és egyéb probléma nem lép fel, akkor ezt helyes adatnak vélheti és „valid" jelzéssel látja el, holott

ez nyilván egy helytelen adat. Nagyméretű adathalmazok esetén az ilyen típusú hibák kiszűrése elég nehéz feladat, hiszen egy jelzés alapján helyes, de gyakorlatilag helytelen adatról van szó.

A műszerek teljes meghibásodása okozhatja azt, hogy egyáltalán nem mérnek adatokat és így hiányzó attribútum értékeket kell kezelni, vagy pedig a saját ellenőrző rendszerük alapján is „invalid", hibás értékeket mérnek. Ezek egyértelműen láthatóak az adathalmazban, azonban a pótlás, javítás sem mindig egyszerű feladat főként ha hosszú távon áll fent egy adott hiba. Különösen a vizsgálthoz hasonló típusú adatbázisok esetén, melyek pl. hőmérséklet, légnedvesség, stb. adatokat tartalmaznak, hiszen az utóbbi időben mi magunk is tapasztalhatjuk, hogy pl. a hőmérséklet egyik napról a másikra akár 15-20 °C-ot is változhat teljesen ad hoc módon. A hagyományos pótlási módszerekkel hosszú távú hibákat, hiányokat, illetve ilyen nagy és gyors ugrásokat nehéz kezelni. Így a bemutatásra kerülő lehetőségek hasznosak lehetnek főként az említett típusú adatokat tartalmazó adatbázisok esetén egy-egy szenzor meghibásodásából eredő hibák, hiányosságok felfedésére, javítására.

Mérési hibák kezelése

Az adatbányászat osztályozó eljárásainak lényege, hogy egy paraméter értékeit jósolni tudjuk a többi paraméter függvényében. Ez alkalmas arra is, hogy egy-egy mérőműszer meghibásodásából fakadó adathibákat felderítsünk és javítsunk. Ennek feltétele, hogy a többi paraméter rendelkezésünkre álljon és lehetőleg minél kevesebb hibát tartalmazzon (élhetünk avval a feltételezéssel, hogy az eleve hibás adatokat már a szenzor kiszűri és nem jellemzőek tömegesen a korábban már említett rejtett mérési hibák). Fontos megjegyezni, hogy az osztályozó modellek pontossága soha nem lesz 100%, a cél az lehet, hogy megpróbáljuk minél hatékonyabban és jobban megközelíteni ezt az ideális állapotot. Így egy lényeges eleme a vizsgálatnak annak meghatározása, hogy melyik modell ad hatékonyan, futási időben is elfogadható eredményt. A feladat megvalósítására az alkalmazott szoftverben a Naive Bayes, Döntési fa algoritmusok lehetnek alkalmasak.

A Naive Bayes algoritmus alapvetően diszkrét adattípusokra alkalmazható, folytonos adatok esetén az algoritmus automatikusan diszkretizálja az attribútum értékkészletét és a jósolt érték az adott kosár közepe. Ez azonban általában nem megfelelő, mert a kosarak középértéke messze esik a tényleges értéktől, azaz túl szélesek a kosarak.

A Döntési fa folytonos adatsorokra is alkalmazható, azonban az adott szoftver kapcsán azt tapasztaltuk több attribútum esetén is (pl. talajhőmérsékletek), hogy ugyanolyan input paraméterezéssel, mint a Naive Bayes esetében a döntési fa algoritmus így nem fut le túl sok folytonos input paraméter okán. Így az egyik megoldás az input paraméterek számának csökkentése, azonban ennek egyik fő hátránya, hogy nem megfelelő attribútumokat elhagyva a jóslás minősége számottevően romlik. A másik lehetőség, hogy a jóslandó paraméter értékeit diszkretizáljuk, így diszkrét adatokra alkalmazzuk az eljárást.

A bevezetésben már említett időszak adatait használjuk a modellek betanítására és a 2008.08.01-2008.12.31 között rendelkezésre álló adatokra alkalmazzuk azokat. A konkrét eredmények bemutatására a közvetlenül a felszínen (0 cm) mért talajhőmérséklet attribútumot (th0) használjuk, mert az értékkészlete elég széles, eleve tartalmaz „invalid" jelzésű adatokat, és ezen attribútum kapcsán sikerült felderíteni rejtett mérési hibákat is az általunk javasolt módszerrel. A korábban már említett probléma itt is fennáll, azaz a Döntési fa algoritmus ezen attribútum kapcsán sem futott le. Hosszas kísérletezések árán

sikerült olyan input paramétereket kiválasztani, melyekkel már jó eredményekkel működött az algoritmus, bár a futási ideje 15-20 szorosa az általunk javasolt lehetőségnek. Ugyanakkor fontos még egyszer megemlíteni, hogyha rossz input paramétereket választunk a modellben, akkor a jóslás minősége teljességgel elfogadhatatlanná válik.

Ennek kiküszöbölésére adódott az ötlet, hogy a jóslandó folytonos attribútum (pl. th0) értékkészletét eleve diszkretizáljuk, kosarazzuk az alkalmazott szoftver nevezett kalkulációinak alkalmazásával (nem közvetlenül az adatbázisban diszkretizálunk). Így már a Döntési fa algoritmus is minden gond nélkül lefut, ráadásul a futási ideje elenyésző a diszkretizálás nélküli esethez képest és a kapott eredmények is jónak tekinthetőek. Milyen méretű kosarakat használjunk, mik legyenek az input paraméterek? Ennek eldöntéséhez szükséges, hogy a különböző paraméterű modelleket összehasonlíthassuk. A felépített modellek összehasonlítására lehetőségünk van az adott szoftverben is (*2. kép*), azonban, mint látni fogjuk evvel kapcsolatban folytonos adatok esetén több probléma is felmerül.

2. kép

Modellek jóságának összehasonlítása

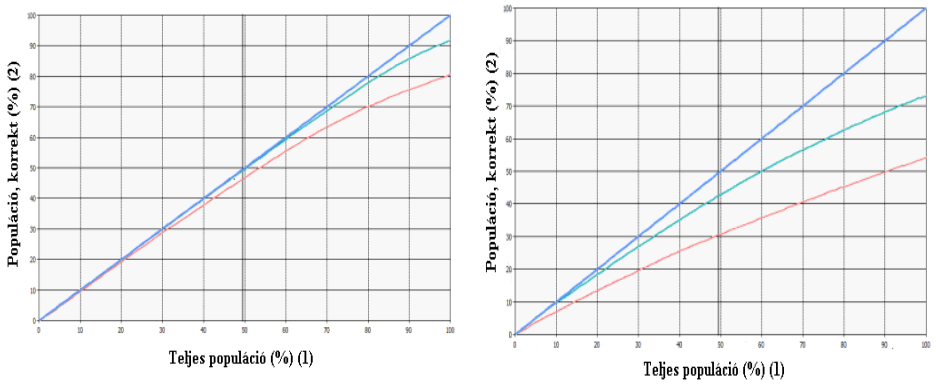


Figure 2. The goodness of modells

The complete population (%)(1), The correct population (%)(2)

A bal oldali diagram a th0-ra (0 cm-en mért talajhőmérséklet) automatikus kosarazást alkalmazó Naive Bayes (piros vonal) és Döntési fa (zöld vonal) algoritmus jóságát mutatja, a jobb oldali pedig az általunk létrehozott 2 szélességű kosarakra diszkretizált th0 modell eredménye. A kapott grafikonok vízszintes tengelyén %-os arányban a populáció mérete látható, míg az y tengely az adott populáció mérethez képest a helyes jóslások arányát mutatja szintén % formátumban. A diagram átlójában elhelyezkedő kék vonal mutatja az ideális esetet, amikor az adott populációra vonatkozó jóslás teljes egészében helyes. A módszerünk jóságát jelző két görbe (piros és zöld) minél közelebb helyezkedik el az ideális állapotot jelző átlóhoz, annál jobb a vizsgált eljárás. Pusztán a grafikonok alapján azt mondhatnánk, hogy az automatikus kosarazás láthatóan jobb eredményt ad. Hiszen pl. a populáció 80%-a esetén az automatikus kosarazás jósága

döntési fára közel 78%, míg Naive Bayes esetén 70%. Ugyanezek az értékek az általunk alkalmazott 2 szélességű kosarak esetén döntési fára körülbelül 63%, míg Naiv eBayes esetén nagyjából 45%-os értéket ér el. Azonban ne felejtsük el, hogy az automatikus kosarak (ezen esetben kb. 6 egység szélesek) és egy így adott helyes jóslás rosszabb lehet, mint a 2 szélességű kosarazás esetén egy éppen rossz jóslás. Az mindkét grafikon esetén látható, hogy a Döntési fa (piros görbe) ad jobb eredményeket a Naive Bayes-szel (zöld görbe) szemben.

A modellek pontosságának számszerű jellemzésére több jól bevált módszer is van (*Abonyi*, 2006; *Bodon*, 2008), ezek általában a jól és a hibásan osztályozott adatok arányából adnak egy jósági értéket, azonban folytonos adatsorok esetén ez megint nehézségekbe ütközik. Így esetünkben hasznos lehet, ha nem csak azt figyeljük, hogy mely diszkrét jóslások helyesek és melyek helytelenek, hanem azt is vegyük figyelembe, hogy a helytelen jóslások mennyire helytelenek. Definiáljuk minden jósolt adatsorra a $d_i^k = \left| x_{mért} - x_{jósolt} \right|$; i=1, … , n, értékeket melyek az adott jósolt érték (attribútum k) abszolút eltérése a tényleges értéktől, majd ezekből egy $D^k = \sum_{i=1}^{n} d_i^k$ értéket, ahol n a rekordok száma. Ezt a mérőszámot kiegészítve avval, hogy mennyi az adott eltérésnél nagyobb eltérést mutató esetek száma a jóslás során egy jó indikátor adódik a diszkretizált folytonos attribútum jósolt értékeinek jóságára nézve. Alkalmazva az itt leírtakat a mi példánkban a jóslások pontosságát tekintve az alábbiak adódnak (*4. táblázat*).

4. táblázat

A különböző modellek pontosságának jellemzése

	folytonos th0	**diszk., 0,5 szélesség**	**diszk., 1 szélesség**	**diszk., 1 szélesség***	**diszk., 2 szélesség**	**rossz input választás**	**automatikus kosarazás**
D^k(1)	27630,4	88908,59	74674,12	106860,9	80618,92	363912,48	146992,85
>1♦	4063	32236	24835	39975	28199	85257	62056
>2♦	998	9621	5632	14221	5300	65609	24207
>3♦	367	3357	1432	4891	940	49570	8015
>4♦	184	788	360	1636	189	35464	2572

th0: a talajszinten mért talajhőmérsékletet (C°) (*groundtemperature (C°) in 0 cm*); folytonos th0: a Naive Bayes algoritmus eredménye, diszkretizálás nélkül (*result of Naive Bayes algorithm without discretization*); diszk., 0,5, diszk., 1 és diszk., 2 szélesség: diszkretizálás 0,5 és 1 illetve 2 szélességű kosarakra (*discretization 0,5, 1 and 2 C° width baskets*); diszk., 1 szélesség*: diszkretizálás 1 szélességű kosarakra ugyanolyan input attribútumokkal, mint folytonos th0 esetén (*discreatization 1 C°width baskets using input like „folytonos th0"*); rossz input választás: diszkretizálás nélkül rossz input választással (*bad input choice without discretization*); automatikus kosarazás: szoftver automatikus kosarazásának alkalmazása (*automatically baskets*).

♦ A sorok az adott hibánál nagyobb eltérést adó jósolt rekordok számát tartalmazzák (*The rows contain the number of predicted records, where the deflection is bigger than the given value*)

Table 4. The accuracy of modells

Summarized absolute average error(1)

Az előbbiek alapján ezért az 1 szélességű kosarak és a th0 attribútum esetén a Döntési fa algoritmus mellett döntöttünk. Ugyan találtunk végül egy olyan input paraméter választást is, amikor a diszkretizálás nélküli jóslás jó eredményeket ad (folytonos th0), azonban ezen paraméterhalmaz meghatározása rendkívül sok időt vesz igénybe és gyakorlatilag a szerencsén is múlik; valamint ezen paraméterezés mellett a futási idő már egy ilyen méretű adatbázis esetén is jelentősen megnövekszik (40 perc a mi példánkban). Másrészt ha rosszul választunk input paraméter halmazt nagyon rossz eredmények is adódhatnak. Ezzel szemben az 1 szélességű kosarakra történő diszkretizálás még megfelelő eredményt ad és két előnye is van: egyszerű az input paraméterválasztás (gyakorlatilag minden releváns paramétert használtunk), valamint a futási ideje nagyságrendekkel kisebb (a mi esetünkben kb. 2 perc, ami 20-ad része a másiknak).

A módszert alkalmazva a th0 attribútumra a 2008.08.01-et követő időszakra a több mint 2,5 hónapon keresztül „invalid" jelzésű adatokat javítottuk, valamint a kapott értékeket összevetettük a „valid" jelzésű mért értékekkel. Több esetben is jelentős eltérés adódott a jósolt és a mért érték között (az, hogy mit tekintünk jelentős eltérésnek az adott paraméter értéktartományától függ), így itt vagy rejtett hibával állunk szemben, vagy a jóslás adott rossz eredményt. Azonban ez utóbbi lehetőségét minimálisra csökkenthetjük azáltal, hogy megpróbáljuk a legjobb megbízhatóságú modellt alkalmazni, ami nem vét nagy hibákat. Másrészt a kapott jósolt eredmény helyességét, helytelenségét az adatbázis egyéb paramétereivel (pl. dátum, léghőmérséklet) összevetve is vizsgálhatjuk. A módszer alapvető eredménye, hogy megmutatja azokat a rekordokat (*5. táblázat*), ahol felmerülhet a rejtett műszerhiba lehetősége.

5. táblázat

Rejtett műszerhiba felderítése és adatjavítás

Dátum (1)	Idő (2)	Status	Eredeti th0 érték (3)	Jósolt th0 érték (4)
2008.08.06	00:00:08	VALID	-0,00099	15,5
2008.08.06	00:10:08	VALID	0,018571	14,5
2008.08.06	00:20:08	VALID	-0,00099	15,5
2008.08.06	00:30:08	VALID	-0,02055	15,5
2008.08.06	00:40:08	VALID	0,018566	15,5

Th0: 0 centiméteren mért talajhőmérséklet (C°) (*ground temperature (C°) in 0 cm*)

Table 5. Explorating crypting instrumental errors

Date(1), Time(2), Origonal th0 values(3), Predicted th0 values(4)

Ezekben az esetekben egyértelműen eldönthető a dátum és a tendenciák figyelembevételével, hogy a „valid" jelzés ellenére a mért adatok hibásak, és így érdemes azokat pótolni a jósolt adatokkal. Azt tapasztaltuk, hogy 2008.08. hónapban a talajhőmérséklet adatok vonatkozásában, több alkalommal is „valid" jelzést kaptak valójában hibás adatok, melyeket egy ekkora adatbázisból nehéz egyéb módon kiszűrni. A módszert alkalmaztuk a vizsgált időszakban hiányzó, illetve hibás talajhőmérséklet adatok pótlására, javítására, valamint felfedtünk több a fentihez hasonló rejtett adathibát is.

ÖSSZEFOGLALÁS

Megmutattuk, hogy az adatbányászat alkalmazásával kinyerhetőek az adatbázisból olyan kapcsolatok, összefüggések, melyeket a hagyományosnak tekintett statisztikai, matematikai módszerekkel nem vagy csak nehezen lehet előállítani. Továbbá bemutattuk annak lehetőségét, hogy adatbányászati módszerekkel hogyan lehet adathibákat kezelni, akár hosszabb távon is, valamint rejtett műszer meghibásodásokból adódó mérési hibákat felfedni és javítani.

IRODALOM

Abonyi J. (2006): Adatbányászat a hatékonyság eszköze, Computerbooks : Budapest 185-236 p.

Bodon F. (2008): Adatbányászat (elektronikus kézirat), http://www.cs.bme.hu/~bodon/magyar/index.html

Han J., Camber M. (2006): Data Mining, Concepts and Techniques - second edition, Morgan Kaufmann Publishers, 772 p.

Vig P. (1995): Éghajlattan, Egyetemi jegyzet, Sopron, 152 p.

Levelezési cím (*Corresponding author*):

Pödör Zoltán
Nyugat-magyarországi Egyetem, Erdőmérnöki Kar, Matematikai Intézet
9400 Sopron, Ady Endre út 5
University of West Hungary, Faculty of Forestry, Mathematics Institute
H-9400 Sopron, Ady Endre út 5.
Tel: +36-99-518-177
e-mail:podzol@emk.nyme.hu