Ivans Samusenko
Daugavpils Universitāte, Latvija

LĒMUMU KOKU KONSTRUĒŠANAS ALGORITMA PIELIETOJUMS SLIMĪBU RISKU ATKLĀŠANAI

Abstract

Use of decision trees building algorithm for risks definition of diseases

Modern medicine is widely using newest technologies and techniques. All of them have a one purpose: to diagnose and treat the diseases. Sometimes, when analysis database is filled, it is difficult to integrate and maintain the main results. This may require the use of artificial intelligence methods, for example - the decision tree construction algorithms. Decision tree construction method is both: a powerful analysis tool and an excellent complement to other data mining tools. Combining such features as easy usage, quick learning process and results testing possibility, the algorithm can be used as a disease diagnosis, prediction of disease risk and detection enhancement. Decision tree construction method allows to get intuitive results presentation and the rules composed in natural language. This advantage makes it possible to use a method not only by experienced users of data mining process, but also to other not professional users, and this significantly extends the usability of method.

Atslēgas vārdi: mākslīgais intelekts, lēmumi koki, C4.5

Ievads

Mūsdienu medicīnā plaši izmanto jaunākās tehnoloģijas un metodes. Tām visām ir mērķis slimību diagnosticēt un to ārstēt. Dažkārt, kad ir apkopota analīžu datu bāze, ir grūti apvienot un saglābāt to galveno rezultātu. Šajā gadījumā var izmantot mākslīgā intelekta metodes, piemēram – lēmumu koku konstruēšanas algoritmus. Lēmumu koku konstruēšanas metode ir gan pats par sevi spēcīgs analizēšanas instruments, gan lielisks papildinājums citiem datu analizēšanas rīkiem. Apvienojot tādas īpašības kā - vienkāršs pielietošanas veids, ātrs apmācības process un rezultāta testēšanas iespēja, šo algoritmu var izmantot gan kā slimības diagnosticēšanas, gan prognozēšanas un slimības risku atklāšanas papildierīci. Lēmumu koku konstruēšanas metode ļauj iegūt intuitīvi saprotamu rezultātu atspoguļojumu un arī dabiskajā valodā formulētus likumus. Šī priekšrocība dod iespēju izmantot metodi ne tikai pieredzējušiem lietotājiem datu analizēšanas jomā, bet arī lietotājiem neprofesionāļiem, kas būtiski paplašina metodes izmantojamību.

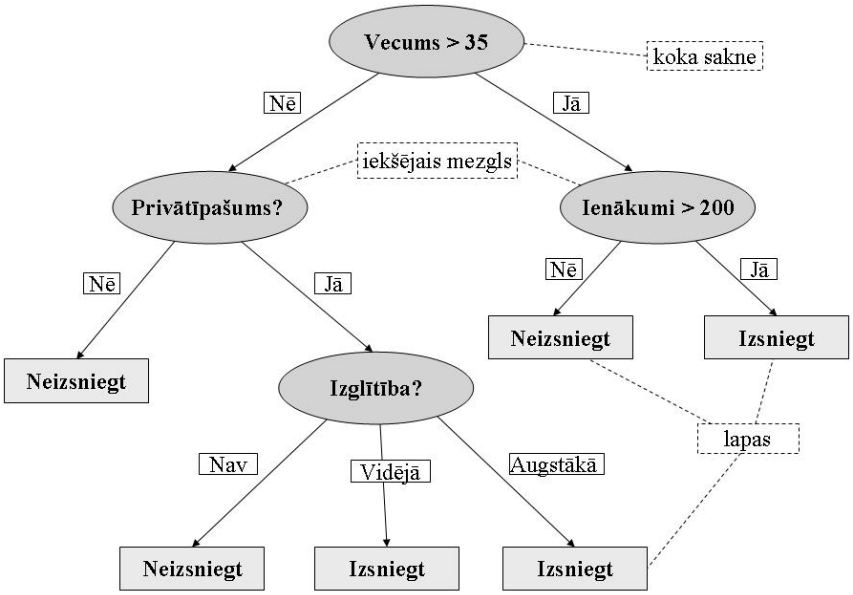
Datu analizēšanas problēmas un to risinājums

Dažkārt statistiskas metodes nedod iespēju cilvēkam iegūt acīmredzamu un saprotamu rezultātu. Viena no svarīgākajām lēmumu koku priekšrocībām ir intuitīvums. Klasificēšanas modelis (Чубукова), kas ir attēlots lēmuma koka veidā, ir vienkārši interpretējams lietotājam.

Piemēram, vienas no statistikas metodēm - faktoru analīzes pielietošanas rezultāti dažkārt nav saprotami un acīmredzami pat pieredzējušam lietotājam.

Lēmumu koka modeļa intuitīvums ir tikai viena no priekšrocībām. Lēmumu koki ļauj iegūt no datu kopas lēmumu pieņemšanas likumus saprotamā valodā. Piemēram: „Ja skats ārā ir saulains un vējš ir vājš, tad spēlēt tenisu".

Lēmumu koka konstruēšanas laikā var atklāt svarīgākos faktorus, kuri ietekmē lēmuma pieņemšanu. Lēmumu koks dod iespēju saprast, kurš faktors ir svarīgāks nekā citi, jo tuvāks lēmumu koka iekšējais mezgls koka saknei, jo svarīgāks tas ir.



1. Zīmējums Lēmumu koka piemērs (Чубукова)

Vēl viena būtiska lēmumu koka priekšrocība ir ātrs apmācības process un testēšanas iespēja. Piemēram, tādas mākslīgā intelekta metodes kā neironu tīkli, ar kuru palīdzību var risināt klasificēšanas uzdevumus, apmācības laiks ir vairākas reizes ilgāks nekā lēmumu kokiem.

Lēmumu koku konstruēšanas ierobežojumi

Lai pielietotu lēmumu koku konstruēšanas algoritmu, jāņem vērā dažus ierobežojumus un rekomendācijas.

Ja nepieskaramies tēmai par datu pirmapstrādi (Арустамов), tad izmantojot lēmumu koku konstruēšanas metodi, var sastapties ar tādu problēmu kā izvēlēties vajadzīgos atribūtus lēmumu koka konstruēšanai. Dažreiz atribūtu atlase nav tik acīmredzama un prognozēt, kurus no tiem ir jāizvēlas lēmumu koka konstruēšanai, nav iespējams. Viens veids kā rīkoties ir - izvēlēties patvaļīgi vienu atribūtu kā klasi (mērķa atribūtu) un visus pārējos par ieejošajiem atribūtiem. Taču iegūtais rezultāts apspoguļos tikai kaut kādu kopēju problēmu, bet ne viņas būtību.

Racionālāks šīs problēmas risinājuma veids ir pieredzējuša lietotāja rekomendācija atribūtu atlasei vai arī - meklēt datu kopā jaunas likumsakarības starp atribūtiem, izmantojot Hi-kvadrāta metodi. Ar šo metodi iespējams uzzināt vai eksistē sakarība starp mērķa un pārējiem atribūtiem. Pamatojoties uz iegūto rezultātu, var konstruēt lēmumu koku, kurš fokusē problēmas vai likumsakarības pētīšanu klases apgabalā.

Nākošā rekomendācija ir apmācības kopas izmērs. Tā ir ļoti būtiska prasība lēmumu koka konstruēšanā. Dažādos literatūras avotos norādīti dažādi skaitļi, taču var apgalvot, ka, lai konstruētu uzticamu lēmumu koku, nepieciešama apmācības kopa ar vismaz 500 piemēriem/ierakstiem. Protams, eksistē tādi uzdevumi, kuros šis skaitlis ir mazāks, piemēram, „Izpētīt pacientus ar kādu retu slimību vai diagnozi". Tādā gadījumā šī rekomendācija nav jāņem vērā.

Lēmumu koka konstruēšana

Lēmumu koku konstruēšana sastāv no sekojošiem etapiem, kuri izpildās rekursīvi, kādēļ nav atrasta pēdējā koka lapa:

1. Izmantojot sašķelšanas kritēriju, atrast labāko atribūtu;
2. Atrast iegūtā mezgla zarus (atribūta vērtības);
3. Atrast dotajā datu bāzē piemērus, kuri apmierina katrā zarā iegūto apgalvojumu;
4. Katram zaram atrast savu labāko atribūtu (tie atribūti, kuri ir jau izmantoti citos mezglos, nav iekļauti meklēšanas kopā);
5. Atkārtot 2 – 4 punktus, kamēr katrā atrastajā zarā visi piemēri pieder vienai klasei vai - vairs nav atribūtu, lai tālāk sašķeltu mezglu (tādā gadījumā par koka lapu jāizvēlas klase, kurai pieder vairāk piemēru, un procentuāli jānorāda piemēru daudzums mezglā).

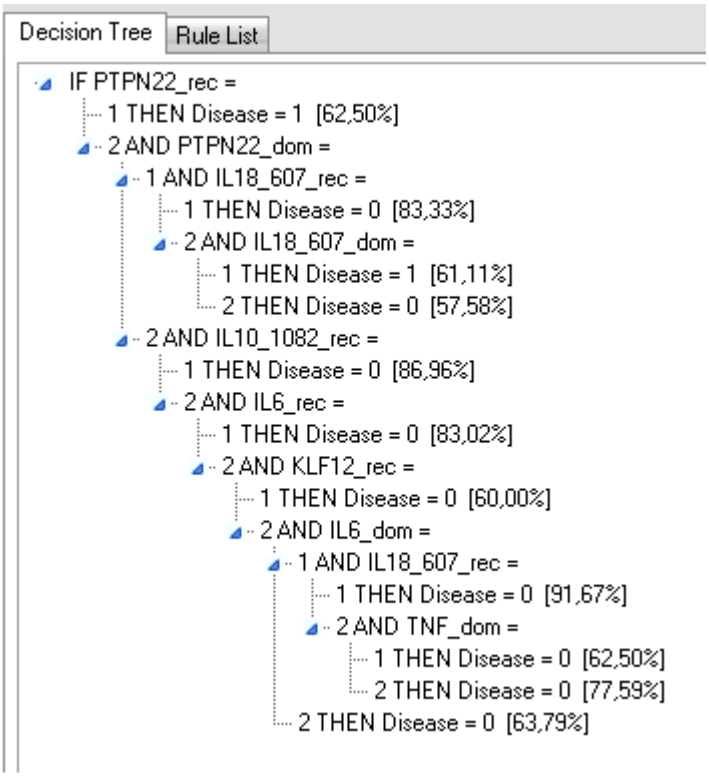
Datu analīze

Ģenētisku analīžu rezultātu analizēšanai tiek izveidota datorprogramma, kurā tiek realizēts lēmumu koku konstruēšanas algoritms C4.5 (Шахиди). Datorprogrammai ir šādas iespējas:

- izlaistu datu pirmapstrāde;
- datu savstarpēju attiecību atklāšana;
- izvelēties lēmumu koka konstruēšanas uzticamību;
- izvēlēties nepieciešamo piemēru skaitu mezglā tā sašķelšanai;
- iegūtā lēmumu koka testēšanas iespēja.

Analizēšanas laikā tiek izmantotas datorprogrammas divas iespējas kā konstruēt lēmumu koku. Pirmā iespēja - intuitīvi izvēlēties nepieciešamo atribūtu skaitu un vēlamo klasi, otrā iespēja – izmantot datorprogrammas datu savstarpējo attiecību atklāšanas ierīci (šī iespēja tiek realizēta ar Hi-kvadrāta metodes palīdzību), kura apspoguļo visas eksistējošās kombinācijas starp atribūtiem un klasēm. Pirmo pieeju ieteicams izmantot, lai noskaidrotu savas hipotēzes. Otro - lai izskatītu datorprogrammas atrastās eksistējošās hipotēzes.

Iegūtos lēmumu kokus var interpretēt kā klasificēšanas un prognozēšanas modeli, tas nozīmē, ka to var izmantot gan kā diagnosticēšanas, gan kā profilakses instrukciju. Protams, vienas slimības rašanos vai gaitu ietekmē daudz faktoru un lēmumu koks cenšas atspoguļot tikai spēcīgākos/spilgtākos faktorus (izņemot gadījumu, kad ir par maz datu vai tiek izmantoti lēmumu koka konstruēšanas uzstādījumi).



2. Zīmējums Lēmumu koks *Disease*

No dotā lēmumu koka (2.zīm.) ir redzams, ka ir tieša saistība starp cilvēka gēniem un viņa veselības stāvokli. Tomēr, lai labāk analizētu iegūto grafisko rezultātu, vēlams interpretēt lēmumu koku likumu veidā.

Uzstādījuma nosaukums	Uzstādījuma vērtība
Piemēru skaits	347
Uzticamība (%)	80
Piemēru skaits mezglā	60
Likumu skaits	10
Atpazīto piemēru skaits	252
Koka testēšana	0 / 0

1. Tabula Lēmumu koka *Disease* uzstādījumi

Likuma Nr.	Likums	Iegūta uzticamība (%)	Piemēru skaits
1.	IF PTPN22_gt = 11 THEN Disease = 1	62,50	5
2.	IF PTPN22_gt = 12 AND IL18_607_gt = 11 THEN Disease = 0	83,33	10
3.	IF PTPN22_gt = 12 AND IL18_607_gt = 12 THEN Disease = 1	61,11	22
4.	IF PTPN22_gt = 12 AND IL18_607_gt = 22 THEN Disease = 0	57,58	19
5.	IF PTPN22_gt = 22 AND IL6_gt = 11 THEN Disease = 0	84,38	54
6.	IF PTPN22_gt = 22 AND IL6_gt = 12 AND IL10_1082_gt = 11 THEN Disease = 0	89,47	17
7.	IF PTPN22_gt = 22 AND IL6_gt = 12 AND IL10_1082_gt = 12 AND TNF_gt = 12 THEN Disease = 0	63,64	7
8.	IF PTPN22_gt = 22 AND IL6_gt = 12 AND IL10_1082_gt = 12 AND TNF_gt = 22 THEN Disease = 0	77,78	35
9.	IF PTPN22_gt = 22 AND IL6_gt = 12 AND IL10_1082_gt = 22 THEN Disease = 0	73,17	30
10.	IF PTPN22_gt = 22 AND IL6_gt = 22 AND IL10_819_gt = 11 THEN Disease = 0	83,33	5
11.	IF PTPN22_gt = 22 AND IL6_gt = 22 AND IL10_819_gt = 12 THEN Disease = 0	56,00	14
12.	IF PTPN22_gt = 22 AND IL6_gt = 22 AND IL10_819_gt = 22 THEN Disease = 0	72,34	34

2. Tabula Lēmumu koka *Disease* likumi

Secinājumi

- Lēmumu koku konstruēšanas metode ir spēcīgs analizēšanas instruments un lielisks papildinājums citiem datu analizēšanas rīkiem;
- Lēmumu koku algoritmus var izmantot gan kā slimības diagnosticēšanas, gan prognozēšanas un slimības risku atklāšanas papildierīci;
- Iegūtos likumus var izmantot kā ekspertsistēmas zināšanas bāzi.

Bibliogrāfija

1. Арустамов, А. Предобработка и очистка данных перед загрузкой в хранилище. http://www.basegroup.ru/library/dw_olap/dataclearing/ [2011.05.05].
2. Чубукова, И. Методы классификации и прогнозирования. Деревья решений. http://www.intuit.ru/department/database/datamining/9/2.html [2011.05.05].
3. Чубукова, И. Задачи Data Mining. Классификация и кластеризация. http://www.intuit.ru/department/database/datamining/5/ [2011.05.05].
4. Шахиди, А. Деревья решений - C4.5 математический аппарат. Часть 1. http://www.basegroup.ru/library/analysis/tree/math_c45_part1/ [2011.05.05].
5. Шахиди, А. Деревья решений - C4.5 математический аппарат. Часть 2 http://www.basegroup.ru/library/analysis/tree/math_c45_part2/ [2011.05.05].