



RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Lietišķo datorsistēmu institūts

Gints JĒKABSONS
Datorsistēmu doktora programmas doktorants

**HEIRISTISKAS METODES DAUDZDIMENSIJU
REGRESIJAS MODEĻU BŪVĒŠANĀ**

Promocijas darba kopsavilkums

Zinātniskais vadītājs
Dr.cs.eng., profesors
J.LAVENDELS

Rīga 2008

UDK 004.85.023(043.2)

Jē 304 h

Jēkabsons G. Heiristiskas metodes daudzdimensiju regresijas modeļu būvēšanā. Promocijas darba kopsavilkums. – Rīga: RTU, 2008. – 38 lpp.

Iespiests saskaņā ar RTU Lietišķo datorsistēmu institūta 2008.gada 22. septembra lēmumu, protokols Nr. 61.

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu Nacionālās programmas „Atbalsts doktorantūras programmu īstenošanai un pēcdoktorantūras pētījumiem” projekta „Atbalsts RTU doktorantūras attīstībai” ietvaros.

ISBN 978-9984-32-855-3

**PROMOCIJAS DARBS
IZVIRZĪTS RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ
INŽENIERZINĀTŅU (datorsistēmu)
DOKTORA GRĀDA IEGŪŠANAI**

Promocijas darbs inženierzinātņu (datorsistēmu) doktora grāda iegūšanai tiek publiski aizstāvēts 2008. gada 22. decembrī plkst. 14.30 Rīgas Tehniskās universitātes Datorzinātņu un informācijas tehnoloģijas fakultātē, Meža ielā 1/3, 202. auditorijā.

OFICIĀLIE RECENZENTI:

Dr.habil.sc.comp., prof. Arkādijs Borisovs
Rīgas Tehniskā universitāte, Latvija

Dr.math., asoc. prof. Māra Gulbe
Latvijas Universitāte, Latvija

Dr.sc.ing., asoc. prof. Dale Dzemydiene
Mikolas Romeris Universitāte, Lietuva

APSTIPRINĀJUMS

Apstiprinu, ka esmu izstrādājis šo promocijas darbu, kas iesniegts izskatīšanai Rīgas Tehniskajā universitātē inženierzinātņu doktora grāda iegūšanai. Promocijas darbs nav iesniegts nevienā citā universitātē zinātniskā grāda iegūšanai.

Gints Jēkabsons.....(paraksts)

Datums:

Promocijas darbs ir uzrakstīts latviešu valodā, satur ievadu, 5 nodaļas, secinājumus, literatūras sarakstu (321 nosaukums), 3 pielikumus, 32 attēlus, 22 tabulas, kopā 186 lappuses.

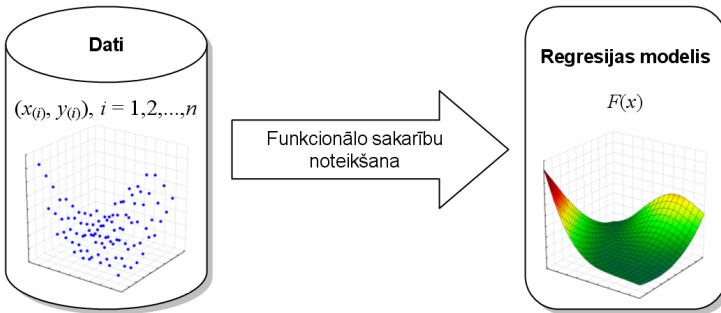
SATURS

VISPĀRĒJS DARBA RAKSTUROJUMS.....	5
Tēmas aktualitāte.....	5
Promocijas darba mērķis.....	6
Darba uzdevumi.....	6
Pētījuma metodes.....	7
Zinātniskie jaunieguvumi un teorētiskās atziņas.....	7
Darba praktiskā vērtība.....	8
Darba aprobācija.....	9
Publikācijas.....	10
ĪSS DARBA SATURS.....	13
Pirmā nodaļa.....	13
Otrā nodaļa.....	15
Trešā nodaļa.....	19
Ceturtnā nodaļa.....	20
Piektā nodaļa.....	25
Pielikumi.....	32
GALVENIE DARBA SECINĀJUMI.....	33
LITERATŪRA.....	35

VISPĀRĒJS DARBA RAKSTUROJUMS

Tēmas aktualitāte. Promocijas darbs ir veltīts daudzdimensiju jeb daudzfaktoru regresijas (turpmāk vienkārši – regresijas) modeļu būvēšanas problēmai, par pamatu izvēloties polinomu regresiju.

Regresijas uzdevums ir definējams kā funkcionālo sakarību pētīšana starp vektorialu ieeju x (sauktu arī par faktoriālo pazīmju vektoru jeb faktoru vektoru) un skalāru kvantitatīvu¹ izeju y (jeb rezultatīvo pazīmi) ar mērķi iegūt modeli šīm sakarībām [Has2001] (skat. 1. attēlu). Regresijas modeļi parasti tiek iegūti induktīvā procesā, izmantojot datus esošos nezināmās funkcionālās sakarības ieeju-izeju (x, y) pārus jeb piemērus. Šādu modeļu iegūšana ir ļoti svarīga, jo tā ļauj izprast un prognozēt pētāmo ar šiem datiem aprakstīto sistēmas uzvedību, kā arī vienkāršā veidā eksperimentēt, manipulējot ar modeļiem, kad manipulēt ar reālo sistēmu var būt neiespējami vai pārāk dārgi.



1. att. Regresijas modeļa iegūšana (n ir datus esošo piemēru skaits)

Regresijas modelēšanai ir izstrādāta virkne dažādu pieeju – lineārā regresija, regresijas koki, splainu (*spline*) metodes, tuvāko kaimiņu metodes, mākslīgo neironu tīkli, simboliskā (*symbolic*) regresija, atbalsta vektoru mašīnas (*support vector machines*) utt. (skat., piemēram, [Has2001, The2006, Wit2005]). Neskatoties uz šādu dažādību, pielietojumos gan zinātnē, gan industrijā viena no vispopulārākajām ir lineārā regresija [Has2001, Mye2002]. Lineārās regresijas visplašāk lietotie modeļi ir polinomi. Polinomu modeļu plašā lietošana galvenokārt ir saistīta ar to elastību, salīdzinošo mazjūtīgumu pret datus esošo troksni, implementēšanas un lietošanas vienkāršību, matemātisko aprēķinu efektivitāti, kā arī parametru vērtību aprēķināšanai nepieciešamo relatīvi nelielo datu apjomu. Pie tam polinomiem ir arī ļoti labi izpētītas to statistiskās īpašības.

Polinomu regresijas modeļu būvēšanā apakškopas atlasē pieejā pirms labākās bāzes funkciju apakškopas meklēšanas, ir jāveic netriviāls pirmaprādes process – jānodedinē galīga izmēra pilnā bāzes funkciju kopa (piemēram, nodedinējot modeļa maksimālo pakāpi p), kurā esošās bāzes funkcijas tiks izmantotas labākās apakškopas veidošanai. Šāda pilnās bāzes funkciju kopas piemeklēšana daudzos gadījumos var izvērsties sarežģītā un ilgā mēģinājumu un kļūdu procesā, īpaši pie paaugstināta faktoru skaita un situācijās, kad

¹ Ja izejas y iespējamās vērtības ir kvalitatīvas vai kategoriskas nevis kvantitatīvas, tad uzdevums tiek saukts par klasificēšanas uzdevumu un parasti tiek risināts ar atšķirīgiem līdzekļiem [Has2001].

pietiekami labas prognozēšanas spējas sasniegšanai ir nepieciešams augstas kārtas modelis. Bez tam šo izvēli, praktisku apsvērumu dēļ, ietekmē arī modeļa būvēšanas procesam nepieciešamo skaitļošanas resursu apjoms – polinomu apakškopas atlasē, pieaugot faktoru skaitam un pieaugot p , „pilnajā” polinomā esošo bāzes funkciju skaits pieaug eksponenciāli [Has2001, Lin2007, Tod2003, Tod2004] – tātad, pat izmantojot visvienkāršākos heuristiskās pārmeklēšanas algoritmus, arī pārmeklēšanai nepieciešamo skaitļošanas resursu apjoms pieaug eksponenciāli. Tas padara apakškopas atlasē nepieciešamo skaitļošanas resursu apjoma problēmu par aktuālu, pat ņemot vērā mūsdienās pieejamās skaitļošanas jaudas.

Promocijas darbā tiek apskatīta iespēja veikt regresijas modeļu būvēšanu bez nepieciešamības izdarīt pieņēmumus par modeļa ar augstu prognozēšanas spēju iegūšanai nepieciešamajām konkrētajām bāzes funkcijām vai tā kārtu, tā vietā adaptējoties konkrētajiem datiem un veidojot prognozēšanas kļūdas minimizēšanai nepieciešamās bāzes funkcijas heuristiskās pārmeklēšanas gaitā. Šāda iespēja ļautu samazināt subjektīvu minējumu ietekmi uz modelēšanas rezultātiem, samazināt modelēšanas metodes lietotāja iepriekšējās pieredzes nepieciešamību (tajā skaitā samazināt paša lietotāja līdzdalības nepieciešamību modelēšanas procesā, automatizējot to), padarīt modelēšanas procesu elastīgāku, kā arī ievērojami ietaupīt skaitļošanas resursus.

Izstrādātā pieceja var būt noderīga ne tikai polinomu regresijā, bet tajā skaitā arī, darbojoties ar cita veida bāzes funkcijām, nelineārajā regresijā, polinomu neironu tīklos, kā arī klasificēšanas uzdevumos.

Promocijas darba mērķis ir izstrādāt tādu polinomu regresijas modeļu būvēšanas pieeju, kas neprasa pilnās bāzes funkciju kopas vai modeļu maksimālās kārtas definēšanu, tā vietā konstruē prognozēšanas kļūdas minimizēšanai nepieciešamās bāzes funkcijas heuristiskās pārmeklēšanas gaitā, adaptējoties konkrētajiem datiem.

Promocijas darba mērķa sasniegšanai tika izvirzīti un atrisināti šādi **darba uzdevumi**:

1. Izpētīt polinomu regresijas funkcionālo sakarību noteikšanu.
2. Izpētīt un sistematizēt polinomu regresijas modeļu novērtēšanas metodes un kritērijus.
3. Izpētīt apakškopas atlasē pieeju polinomu regresijas modelēšanā:
 - a) izanalizēt un sistematizēt apakškopas atlasē metožu īpašības;
 - b) izanalizēt un sistematizēt regresijas modeļu būvēšanai pielietojamos heuristiskos pārmeklēšanas algoritmus;
 - c) izanalizēt un sistematizēt apakškopas atlasē pieejai esošās alternatīvas;
 - d) identificēt un izanalizēt apakškopas atlasē pieejas trūkumus.
4. Pamatojoties uz veiktajiem pētījumiem, izstrādāt adaptīvu polinomu regresijas modeļu būvēšanas pieeju, kas neprasa izdarīt pieņēmumus par modelī iekļaut nepieciešamajām konkrētajām bāzes funkcijām vai tā kārtu.
5. Veikt apakškopas atlasē pieejas un izstrādātās pieejas salīdzināšanu, ņemot vērā to raksturīgākās īpašības regresijas modeļu būvēšanā.
6. Novērtēt izstrādātās pieejas efektivitāti praktiskos regresijas modeļu būvēšanas eksperimentos, ņemot vērā ar to iegūto modeļu prognozēšanas kļūdu un to iegūšanai nepieciešamos skaitļošanas resursus un izmantojot reālās pasaules datu kopas, salīdzinot to ar pilno polinomu regresiju, biežāk lietotajām apakškopas atlasē

metodēm, kā arī ar plaši pazīstamām un lietotām regresijas modelēšanas metodēm, kas nesakņojas lineārajā polinomu regresijā.

Pētījuma metodes. Promocijas darba teorētiskie pētījumi balstās uz literatūras pētīšanu, matemātisko statistiku, mašīnprogrammēšanas teoriju, mākslīgā intelekta teoriju, asimptotisko analīzi. Darbā ir izmantotas arī programmatūras inženierijas metodes pētījumu rezultātu implementācijai, kā arī eksperimentālie pētījumi dažādo regresijas modelēšanas metožu salīdzināšanai.

Darbā ir šādi **zinātniskie jaunieguvumi un teorētiskās atziņas**:

1. Regresijas modeļu būvēšanas apakškopas atlases pieeja ir apskatīta kā stāvokļu telpas pārmeklēšanas problēma. Ir izanalizētas tajā pielietojamās heuristiskās pārmeklēšanas metodes, nodalot piecas pamatīpašības, ar kurām ir iespējams pilnībā raksturot katru atsevišķo apakškopas atlases metodi. Ir sistematizētas katras nodalītās īpašības praksē biežāk lietotās variācijas, tai skaitā ir veikts regresijas modeļu būvēšanā biežāk lietoto heuristisko pārmeklēšanas algoritmu apkopojums, analīze un klasifikācija.
2. Ir izstrādāta vispārīga adaptīvas bāzes funkciju konstruēšanas (*Adaptive Basis Function Construction*, ABFC) pieeja regresijas modeļu būvēšanai. Tā var kalpot kā universāls karkass, uz kura pamata ir iespējams izstrādāt dažādas jaunas adaptīvas regresijas modeļu būvēšanas metodes ar šādām (no apakškopas atlases metodēm atšķirīgām) īpašībām:
 - nav nepieciešams definēt pilno bāzes funkciju kopu vai izvēlēties modeļu maksimālo kārtu;
 - ir iespējams ģenerēt neierobežotas sarežģītības un kārtas modeļus;
 - modeļu būvēšanai nepieciešamo bāzes funkciju konstruēšana, un līdz ar to arī vēlamās sarežģītības noteikšana, tiek veikta pārmeklēšanas gaitā;
 - izveidojamo modeļu iespējamo funkcionālo formu paaugstinātās elastības dēļ tiek iegūta paaugstināta spēja piemēroties datos esošām sarežģītām nelineārām daudzfaktoru sakarībām;
 - salīdzinājumā ar apakškopas atlases metodēm, tiek iegūtas samazinātas skaitļošanas resursu prasības.
3. Ir apkopotas un salīdzinātas apakškopas atlases un ABFC raksturīgākās īpašības regresijas modeļu būvēšanā, tajā skaitā no heuristiskās pārmeklēšanas viedokļa.
4. Konstatēts, ka pie fiksēta faktoru skaita stāvokļu telpas stāvokļu zarošanās koeficients apakškopas atlases pieejā palielinās eksponenciāli, bet ABFC pieejā – lineāri. Izdarot pieņēmumu par pārmeklēšanas rezultātā atrastā modeļa bāzes funkciju skaitu, kopējie pārmeklēšanai nepieciešamie skaitļošanas resursi apakškopas atlases pieejā palielinās eksponenciāli (pat izmantojot visvienkāršākos heuristiskās pārmeklēšanas algoritmus), bet ABFC pieejā – polinomiski.
5. Konstatēts, ka, relatīvi pret apakškopas atlases pieejas pārmeklēšanas ātrdarbību, ABFC pieejā pārmeklēšanas ātrdarbība pieaug līdz ar faktoru skaita un izvēlētais maksimālās kārtas palielināšanos. Pie tam, tā kā apakškopas atlases pieejā modeļu maksimālā kārtā ir jāpiemeklē, jo tā iepriekš nav zināma, tad modelēšanas procesam nepieciešamie skaitļošanas resursi var vēl vairākkārtēji palielināties.

6. Ir izstrādāts un implementēts piedāvātās ABFC pieejas speciāls gadījums praktiskai polinomu regresijas modeļu būvēšanai.
7. Ir izstrādāta adaptīvi konstruēto bāzes funkciju modeļu intensīvas pārmeklēšanas izraisītās pārmērīgās piemērošanās un paaugstinātas modeļu izvēles dispersijas samazināšanas metode, kas izmanto šķērsvalidēšanu un modeļu vidējošanu.
8. Izstrādātā modeļu būvēšanas pieeja ir pielietojama ne tikai polinomu regresijā ar nenegatīvu pakāpju bāzes funkcijām, bet arī ar cita veida bāzes funkcijām, polinomu neironu tīklos, ar nelineārās regresijas modelēšanas metodēm, kā arī klasificēšanas problēmu risināšanā.
9. Ir novērtēta izstrādāto metožu efektivitāte polinomu regresijas modeļu būvēšanā, izmantojot reālās pasaules datu kopas, salīdzinot to ar pilno polinomu regresiju, biežāk lietotajām apakškopas atlases metodēm, kā arī ar plaši pazīstamām un lietotām regresijas modelēšanas metodēm, kas nesakņojas lineārajā polinomu regresijā.
10. Darbā iegūtās atziņas un veiktajos empīriskajos eksperimentos iegūtie secinājumi var tikt izmantoti regresijas modelēšanas problēmu pētīšanā dažādās nozarēs, kā arī var kalpot kā metodiski priekšnoteikumi turpmākiem pētījumiem šajā virzienā.

Darba praktiskā vērtība. Darbā ir iegūti šādi praktiskie rezultāti:

1. Ir izstrādāta regresijas modeļu būvēšanas pieeja, kas neprasa lietotājam izdarīt pieņēmumus par modeļa būvēšanai nepieciešamajām konkrētajām bāzes funkcijām vai maksimālo kārtu. Pieveicējot ļauj samazināt intuitīvu minējumu ietekmi uz modelēšanas rezultātiem, samazināt modelēšanas metodes lietotāja iepriekšējās pieredzes nepieciešamību, tajā skaitā samazināt paša lietotāja līdzdalības nepieciešamību modelēšanas procesā, automatizējot to. Tas viss īpaši atvieglo daudzfaktoru regresijas modelēšanas uzdevumu risināšanu pielietojumos, kur pilnībā izprast pētāmo sistēmu un tajā notiekošos procesus ir vai nu neiespējami vai nepraktiski, jo prasa pārāk lielu laiku vai naudas resursus un piepūli.
2. Izmantojot izstrādāto pieeju ir iespējams veikt regresijas modeļu būvēšanu, izmantojot ievērojami mazākus skaitļošanas resursus, nekā tos, kas nepieciešami apakškopas atlases metodēm, lai sasniegtu līdzīgus rezultātus.
3. Izstrādātās regresijas modelēšanas metodes ir implementētas programmatūras modulī, kuru vienkāršā veidā ir iespējams pielietot regresijas modelēšanā dažādās problēmsfērās, kā arī iestrādāt citos programmatūras rīkos. Par izstrādāto programmatūru tika arī publicēts ziņojums Latvijas augsto tehnoloģiju katalogā „High Tech in Latvia”.
4. Piedaloties Eiropas Savienības 6. Ietvara Programmas Mērķorientētajā zinātniskajā projektā „Kompozīto materiālu izmantošanas pilnveidošana drošā lidmašīnu konstrukciju projektēšanas praksē balstoties uz spēju precīzi modelēt konstrukciju sabrukumu” („Improved Material Exploitation at Safe Design of Composite Airframe Structures by Accurate Simulation of Collapse”), COCOMAT, Nr. 6785 (RTU), Nr. AST3-CT-2003-502723 (EK) (projekta vadītājs R. Rikards, 2004.-2008. gads) [Kal2008b], izmantojot izstrādātās regresijas modelēšanas metodes implementāciju, iegūtie regresijas modeļi tika iestrādāti materiālu konstrukciju sabrukšanas procesa ātrās prognozēšanas programmatūras rīkā, kas ir paredzēts

zinātniekiem un inženieriem ātrai, vienkāršai un precīzai dažāda veida ribotu konstrukciju sabrukšanas procesu pētīšanai.

5. Promocijas darbā izstrādātās regresijas modelēšanas metodes ir ieviestas Rīgas Tehniskās universitātes Materiālu un Konstrukciju institūtā kompozītmateriālu konstrukciju uzvedības modelēšanā.
6. Darba izstrādes gaitā veiktās regresijas modelēšanas un heuristiskās pārmeklēšanas literatūras apkopošanas un analīzes rezultāti tika ieviesti mācību procesā Rīgas Tehniskās universitātes Datorzinātņu un informācijas tehnoloģijas fakultātes Lietišķo Datorsistēmu institūta maģistrantūras lekciju kursā.

Darba aprobācija ir notikusi šādās starptautiskās konferencēs un zinātniskos semināros:

- 7th ASMO-UK/ISSMO International Conference on Engineering Design Optimization, Association for Structural and Multidisciplinary Optimization in the UK (ASMO-UK) (2008. gada. 7-8. jūlijs, Bāta, Lielbritānija);
- International Conference on Computer, Electrical, and Systems Science, and Engineering, CESSE 2008, World Academy of Science, Engineering, and Technology, WASET (2008. gada 25.-27. aprīlis, Roma, Itālija);
- International Conference on Welded Structures, DFE 2008 (2008. gada. 24-26. aprīlis, Miškolca, Ungārija);
- Scientific conference “Applied Communication and Information Technologies, AICT” (2008. gada 10.-11. aprīlis, Latvijas Lauksaimniecības universitāte, Jelgava, Latvija);
- RTU 48th International Scientific Conference, apakšsekcija „Applied Computing” (2007. gada 10. oktobris, Rīgas Tehniskā universitāte, Rīga, Latvija);
- 12th International Conference “Mathematical Modelling and Analysis”, MMA2007 (2007. gada 30.maijs - 2. jūnijs, Trakai, Lietuva);
- 2nd international workshop on Surrogate Modelling and Space Mapping for Engineering Optimization, SMSMEO-06 (2006. gada 9.-11. novembris, Dānijas Tehniskā universitāte, Kopenhāgena, Dānija);
- RTU 47th International Scientific Conference, apakšsekcija „Applied Computing” (2006. gada 13. oktobris, Rīgas Tehniskā universitāte, Rīga, Latvija);
- 11th International Conference “Mathematical Modelling and Analysis”, MMA2006 (2006. gada 1.-4. jūnijs, Jūrmala, Latvija);
- International Conference on Electrical and Control Technologies 2006, ECT2006 (2006. gada 5. maijs, Kauņas Tehnoloģijas universitāte, Kauņa, Lietuva);
- IADIS International Conference, Applied Computing 2006 (2006. gada 4. marts, Mondragon universitāte, Sansebastjana, Spānija);
- RTU 46th International Scientific Conference, apakšsekcijas „Applied Computing” un „Boundary Field Problems and Computer Simulation” (2005. gada 13.-14. oktobris, Rīgas Tehniskā universitāte, Rīga, Latvija).

Promocijas darba zinātniskie rezultāti ir pielietoti trīs pētniecības projektos:

- Eiropas Savienības 6. Ietvara Programmas Mērķorientētais zinātniskais projekts (STREP) „Kompozīto materiālu izmantošanas pilnveidošana drošā lidmašīnu konstrukciju projektēšanas praksē balstoties uz spēju precīzi modelēt konstrukciju sabrukumu” („Improved Material Exploitation at Safe Design of Composite

Airframe Structures by Accurate Simulation of Collapse”), COCOMAT, Nr. 6785 (RTU), Nr. AST3-CT-2003-502723 (EK) (RTU puses projekta vadītājs R. Rikards, 2004.-2008. gads) [Kal2008b];

- Latvijas Izglītības un Zinātnes ministrijas un Rīgas Tehniskās universitātes pētniecības projekts R7397 „Ribotu kompozīto konstrukciju nestspējas optimizācija ar eksperimentālu validāciju” (projekta vadītājs K. Kalniņš, 2008. gads);
- Latvijas Izglītības un Zinātnes ministrijas un Rīgas Tehniskās universitātes pētniecības projekts R7193 „Kompozīto konstrukciju nestspējas īpašību skaitliskā modelēšana un eksperimentāla testēšana” (projekta vadītājs K. Kalniņš, 2007. gads).

Publikācijas. Veikto pētījumu rezultāti ir atspoguļoti 23 publikācijās starptautiskos recenzējamos zinātniskos izdevumos:

1. Kalniņš, K., Jēkabsons, G., Beitlers, R., Ozoliņš, O. Optimal design of fiberglass panels with physical validation. Scientific Proceedings of Riga Technical University, Construction Science, ISSN: 1407-7493, Riga, Latvia: RTU, 2009, 13 p. (accepted)
2. Kalnins, K., Ozolins, O., Jekabsons, G., Metamodels in design of GFRP composite stiffened deck structure. Proceedings of 7th ASMO-UK/ISSMO International Conference on Engineering Design Optimization, Association for Structural and Multidisciplinary Optimization in the UK (ASMO-UK), Bath, UK, 2008, 11 p. (in print)
3. Jekabsons, G. Ensembling adaptively constructed polynomial regression models. International Journal of Intelligent Systems and Technologies (IJIST), Vol. 3, No. 2, ISSN: 1305-6417, 2008, pp. 56-61. (<http://www.waset.org/ijist/v3/v3-2-11.pdf>)
4. Jekabsons, G. Ensembling adaptively constructed polynomial regression models. International Conference on Computer, Electrical, and Systems Science, and Engineering, CESSE 2008, Proceedings of World Academy of Science, Engineering, and Technology, WASET, Vol. 28, ISSN: 1307-6884, Rome, Italy, 2008, pp. 162-167.
5. Jekabsons, G., Lavendels, J. A heuristic approach for surrogate modelling of electro-technical systems. Proceedings of International Conference “Electrical and Control Technologies ECT-2008”, ISSN: 1822-5934, Kaunas, Lithuania, 2008, pp. 62-67.
6. Jekabsons, G., Lavendels, J. A heuristic approach for surrogate modelling. Proceedings of Conference “Applied Communication and Information Technologies, AICT”, ISBN: 978-9984784687, Jelgava, Latvia: Latvia University of Agriculture, 2008, pp. 11-20.
7. Kalnins, K., E. Eglitis, G. Jekabsons, R. Rikards. Metamodels for optimum design of laser welded sandwich structures. Proceedings of International Conference on Design, Fabrication, and Economy of Welded Structures 2008, DFE2008, ISBN: 978-1904275282, Miskolc, Hungary, 2008, pp. 119-126.
8. Jekabsons, G., Lavendels, J. Polynomial regression modelling using adaptive construction of basis functions. Proceedings of IADIS International Conference, Applied Computing 2008, ISBN: 978-9728924560, Mondragon unibertsitatea, Algarve, Portugal, 2008, pp. 269-276.
9. Jekabsons, G., Kalnins, K., Eglitis, E. Polynomials in metamodeling of glass fibre bar stability. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 34, ISSN: 1407-7493, Riga, Latvia: RTU, 2008, pp. 150-159.

10. Jekabsons, G., Lavendels, J. An approach for polynomial regression modelling using construction of basis functions. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 34, ISSN: 1407-7493, Riga, Latvia: RTU, 2008, pp. 138-149.
11. Jekabsons, G., Lavendels, J. A Comparison of Heuristic Methods for Polynomial Regression Model Induction. Mathematical modelling and analysis, The Baltic Journal on Mathematical Applications, Numerical Analysis and Differential Equations, A.Buikis (ed.), Vol. 13, No. 1, ISSN print: 1392-6292, ISSN online: 1648-3510, Vilnius, Lithuania: Technika, 2008, pp. 17-27.
12. Jekabsons, G., Lavendels, J. Approximation of economical data using an approach of adaptive polynomial basis function construction. Annual Proceedings of Vidzeme University College "ICTE in Regional Development", ISBN: 9984633101, Valmiera, Latvia: Vidzeme University College, 2007, pp. 8-14.
13. Kalnins, K., Jekabsons, G., Janushevskis, J., Eglitis, E., Rikards, R. Different approximation functions in surrogate modelling of sandwich structures. 2nd international workshop on Surrogate Modelling and Space Mapping for Engineering Optimization (SMSMEO-06), Technical University of Denmark, Denmark, Copenhagen, 2006, Optimization and Engineering, International Multidisciplinary Journal to Promote Optimization Theory & Applications in Engineering Sciences, USA: Springer, 2007, 11 p. (submitted)
14. Jēkabsons, G. Schwarz weights for easy interpretation of the results of regression model comparison. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 30, ISSN: 1407-7493, Riga, Latvia: RTU, 2007, pp. 97-103.
15. Jekabsons, G., Lavendels, J., Sitikov, V. Model evaluation and selection in multiple nonlinear regression analysis. Mathematical modelling and analysis. The Baltic Journal on Mathematical Applications, Numerical Analysis and Differential Equations, Vol. 12, No. 1, ISSN print: 1392-6292, ISSN online: 1648-3510, Vilnius, Lithuania, Technika, 2007, pp. 81-90.
16. Jekabsons, G., Lavendels, J. A heuristic approach for regression model selection of electro-technical systems. Proceedings of International Conference Electrical and Control Technologies 2006, ISBN: 9955250542, Lithuania, Kaunas: Kaunas University of Technology, 2006, pp. 456-461.
17. Jekabsons, G., Lavendels, J. A heuristic approach of model selection in multiple nonlinear regression analysis. Proceedings of IADIS International Conference, Applied Computing 2006, Mondragon unibertsitatea, Spain, San Sebastian, 2006, pp. 524-527.
18. Jēkabsons, G. Heuristics in multiple nonlinear regression analysis. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 26, ISSN: 1407-7493, Riga, Latvia: RTU, 2006, pp. 234-243.
19. Jēkabsons, G., Lavendels, J. Evaluation of model selection criterions in multiple nonlinear regression analysis, Scientific Proceedings of Riga Technical University, Boundary Field Problems and Computer Simulation, Series 5, Vol. 25, ISSN: 1407-7493, Riga, Latvia: RTU, 2006, pp. 67-81.
20. Kalniņš, K., Jēkabsons, G., Skukis, E., Sproģis, R. Metamodels for the optimum design of corrugate load-bearing profile plates. Scientific Proceedings of Riga

- Technical University, Architecture and Construction Science, Series 2, Vol. 6, ISSN: 1407-7493, Riga, Latvia: RTU, 2005, pp. 136-145.
21. Jēkabsons, G., Kalniņš, K. Konstrkciju uzvedības metamodeļa izvēle, izmantojot heiristisku stāvokļu telpas pārmeklēšanu. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 22, ISSN: 1407-7493, Riga, Latvia: RTU, 2005, pp. 266-276.
 22. Jēkabsons, G., Lavendels, J., Kozlova, S. Hipotēzes komplikētības reducēšana daudzfaktoru regresijā. Scientific Proceedings of Riga Technical University, Computer Science, Series 5, Vol. 22, ISSN: 1407-7493, Riga, Latvia, 2005: RTU, pp. 283-294.
 23. Jēkabsons, G., Lavendels, J. Mazāko kvadrātu un galīgo elementu aproksimācijas principu koplietojums regresijas analizē. Scientific Proceedings of Riga Technical University, Boundary Field Problems and Computer Simulation, Series 5, Vol. 22, ISSN: 1407-7493, Riga, Latvia: RTU, 2005, pp. 277-282.

ĪSS DARBA SATURS

Pirmā nodaļa. Nodaļā tiek apskatīta regresijas funkcionālo sakarību noteikšana vispārīgā veidā, lineārā un polinomu regresija, problēmas, kas saistītas ar modeļu nepilnīgu piemērošanos (*underfitting*) un pārmērīgu piemērošanos (*overfitting*) datiem, kā arī trīs visbiežāk lietotās pieejas (un konkrētas to metodes un kritēriji) regresijas modeļu novērtēšanai un salīdzināšanai: statistiskā hipotēžu testēšana, sarežģītības „sodīšanas” (*complexity penalization*) kritēriju izmantošana un validēšanas metožu izmantošana.

Regresijas modelēšanā, lai aprakstītu sakarības dotajos datos, visbiežāk tiek lietots lineārās regresijas modelis [Has2001] – tiek pieņemts, ka aproksimējamā mērķa funkcija ir tās parametros lineāra. Vienkāršākais lineārais modelis ir lineārs gan tā parametros, gan tā faktoros un to var definēt kā faktoru x summu:

$$F(x) = a_0 + \sum_{i=1}^d a_i x_i, \quad (1)$$

kur x_i ir i -tais faktors; d ir faktoru skaits; a ir modeļa parametri, kuru vērtības tiek aprēķinātas minimizējot izvēlēto noviržu kritēriju – vidējo kvadrātisko kļūdu; šī modeļa pirmais terms a_0 tiek saukts arī par brīvo konstanti. Pie tam modeļa parametru skaits parasti nedrīkst pārsniegt apmācības kopas piemēru skaitu.

Vispārīgā veidā lineāru modeli var definēt kā summu no bāzes funkcijām [Has2001]:

$$F(x) = \sum_{i=1}^k a_i f_i(x), \quad (2)$$

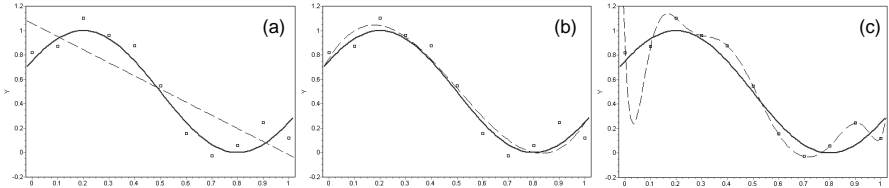
kur $f_i(x)$ ir bāzes funkcijas, kas var būt faktoru x jebkura veida transformācijas un kombinācijas; k ir modelī iekļauto bāzes funkciju skaits (vienāds ar modeļa parametru skaitu). Katra modelī iekļautā bāzes funkcija, kopā ar tās parametru a_i , atbilst vienam polinoma termam. Pie tam modeļa brīvā konstante tiek iegūta, nedefinējot $f_i(x) = 1$.

Polinomu modelēšanas implementēšanas vienkāršība un matemātisko aprēķinu efektivitāte ir lielā mērā saistīta ar to, ka, darbojoties ar lineāriem modeļiem, to parametru vērtību aprēķināšanai nav nepieciešamības pēc laikietilpīgām iteratīvām parametru optimizācijas metodēm, kā tas ir, darbojoties ar nelineāriem modeļiem. Tā vietā parametru vērtības ir iespējams aprēķināt tiešā veidā, izmantojot „parasto” mazāko kvadrātu metodi [Mil2002, Rao1999, Wol2006].

Funkcionālo sakarību modelēšanai var izdalīt divus uzdevumus – izskaidrošana un prognozēšana [Cox1974, Has2001]. Izskaidrošanas uzdevumā ir nepieciešams atrast tādu datiem piemērotu modeli, kas ir maksimāli vienkāršs un interpretējams, atstājot tā prognozēšanas precizitāti otrajā plānā, cenšoties labāk izprast pētāmās sistēmas vai procesa fundamentālo dabu. Turpretī prognozēšanas uzdevumā galvenais mērķis ir modeļa ar maksimālu prognozēšanas spēju atrašana. Prognozēšanas uzdevums ir vairāk formalizējams, un tajā pielietojamās metodes parasti nav atkarīgas no pielietojuma sfēras. Promocijas darbā uzmanība tiek pievērsta tieši prognozēšanas uzdevumam.

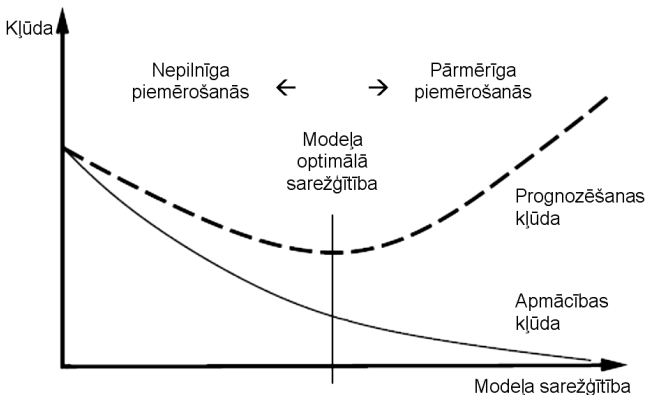
Izvēloties datiem piemērotāko regresijas modeli, vispārīgā gadījumā nav ieteicams ņemt vērā tikai modeļu apmācības kļūdu. Ir jāņem vērā arī modeļu sarežģītība [Dre2006, Has2001], ko lineārajā regresijā parasti mēra ar modeļa parametru skaitu. Pārmērīgi vienkāršs modelis nav spējīgs pietiekami labi aprakstīt apmācības kopā esošās vispārīgās likumsakarības – nepilnīga piemērošanās (2. attēla (a) daļa). Turpretī pārmērīgi sarežģīts

modelis cenšas aprakstīt arī nejauši izveidojušās šķietamās likumsakarības, kas apmācības kopā ir sastopamas tikai tās galīgā rakstura dēļ (bet nebūtu sastopamas, ja apmācības kopā būtu cita piemēru kombinācija vai arī tā būtu vienkārši lielāka), kā arī tur esošo troksni – pārmērīga piemērošanās (2. attēla (c) daļa).



2. att. Trīs dažādu sarežģītību modeļu līknes, kas aproksimē vienus un tos pašus ar punktiem parādītos datus: (a) pirmās kārtas polinoms (nepilnīga piemērošanās); (b) kubiskais polinoms (vispiemērotākais no apskatītajiem modeļiem); (c) desmitās kārtas polinoms (pārmērīga piemērošanās). Patiesā modeļa līkne ir parādīta ar treknu līniju, bet aproksimējošo modeļu līknes – ar tievu raustītu līniju

Tā kā iegūtais regresijas modelis turpmāk tiks izmantots y vērtību prognozēšanai arī pie tādām x vērtībām, kādas apmācības kopā netika dotas, tad mērķis ir atrast modeli nevis ar vismazāko apmācības kļūdu, bet gan ar vismazāko prognozēšanas (jeb vispārināšanas) kļūdu. Apmācības kopas kļūda nedod pietiekamu informāciju par modeļa prognozēšanas spēju. Palielinoties modeļa sarežģītībai, apmācības kopas kļūda gandrīz vienmēr samazināsies. Taču prognozēšanas kļūda pēc minimuma sasniegšanas, tālāk palielinoties modeļa sarežģītībai, atkal sāk palielināties (sk. 3. attēlu). Tas nozīmē, ka ir nepieciešams izvēlēties tādu modeli, kas atbilst labam kompromisam starp modeļa vienkāršību un tā precizitāti apmācības kopā.



3. att. Nepilnīga un pārmērīga piemērošanās, modeļa optimālā sarežģītība (adaptēts no [Has2001])

Polinomu regresijā šī kompromisa meklēšana neattiecas tikai uz polinoma kārtas izvēli. Parasti, lai pietiekami labi aprakstītu datus esošās likumsakarības, ne visas pilnajā polinomā esošās bāzes funkcijas ir nepieciešamas [Egl1981, Had2001, Has2001, Mil2002]. Tāpat arī daļa no x faktoriem var būt nevajadzīgi (pārmērīgi liels trokšņu līmenis, kas neļauj iegūt nekādu aproksimācijai lietderīgu informāciju) vai redundanti (paaugstināta korelācija ar vienu vai vairākiem citiem faktoriem), kas, tāpat kā liekās bāzes funkcijas, ir tikai lieks trokšnis datus [Blu1997, Guy2006, Koh1997, Mil2002, Mol2002]. Tādēļ modelī ir lietderīgi iekļaut tikai visnepieciešamāko bāzes funkciju (un līdz ar to arī faktoru) kombināciju (veidojot „nepilnu” polinomu), tādā veidā samazinot modeļa sarežģītību un uzlabojot tā prognozēšanas spēju [Had2001, Has2001, Mil2002].

Modeļu novērtēšanas problēmu var iedalīt divos veidos, kuriem ir atšķirīgi mērķi [Dre2006, Guy2003, Has2001]:

- modeļu izvēle (*model selection*) – tās mērķis ir novērtēt modeļu-kandidātu prognozēšanas spēju dotajos datos, lai izvēlētos labāko;
- modeļu galējā novērtēšana (*model assessment*) jeb modeļu testēšana – tās mērķis ir aprēķināt modeļu izvēles rezultātā iegūtā modeļa prognozēšanas kļūdas maksimāli tuvinātu novērtējumu, iegūstot informāciju par iegūtā modeļa sagaidāmo kļūdu turpmākajos tā pielietojumos.

Modeļu izvēles problēma parasti ir saistīta ar modeļa struktūras (jeb tajā iekļauto bāzes funkciju kombinācijas) optimizēšanu, kad jāatrod tāda modeļa struktūra, kas vislabāk atbilst kompromisam starp vienkāršību un spēju piemēroties dotajiem datiem – tāpat labāko prognozēšanas spēju. Tādēļ šajā problēmā parasti nav būtiski, vai lietotais novērtējuma mērs dod interpretējamu modeļa prognozēšanas spējas novērtējumu tuvinātas prognozēšanas kļūdas vērtības veidā, vai arī neinterpretējamu relatīvu modeļa novērtējumu.

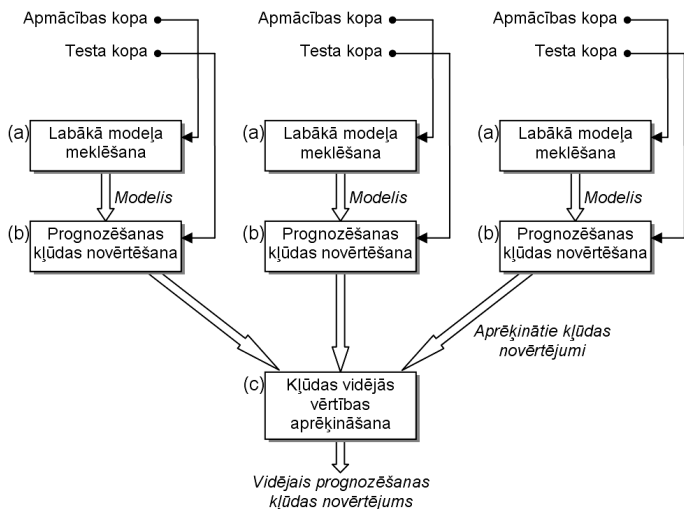
Tā kā modeļu prognozēšanas spēja parasti nav tiešā veidā novērtējama, to ir jānovērtē netiešā veidā. Modeļu izvēles problēmas risināšanas klasiskā pieeja ir statistiskā hipotēžu testēšana. Tomēr tai piemīt virkne īpašību, kuras padara šo pieeju modeļu izvēles problēmas risināšanai mazāk piemērotu. Literatūrā ir piedāvātas arī citas pieejas, kuras galvenokārt var iedalīt divās pamatkategorijās: sarežģītības sodīšanas kritēriju izmantošana un validēšanas metožu izmantošana. Sarežģītības sodīšanas kritēriji parasti ir paredzēti modeļu izvēles problēmas risināšanai, taču validēšanas metodes ir pielietojamas gan modeļu izvēlē, gan galējā novērtēšanā.

Tālāk šajā nodaļā tiek detalizēti apskatītas visas trīs minētās modeļu novērtēšanas pieejas un tām atbilstošās pazīstamākās metodes/kritēriji. Apskatītās pieejas un to priekšrocības un trūkumi attiecas arī uz cita veida regresijas modeļiem un modelēšanas metodēm, ne tikai uz polinomu regresiju, lai gan specifiski mēri un kritēriji var atšķirties.

4. attēlā ir parādīta validēšanas procesa kopējā darbības shēma galējā modeļu novērtēšanā, ja tas sastāv no trīs iterācijām (piemēram, izmantojot 3 daļu šķērsvālidēšanu), kur katrā iterācijā apmācības kopa un testa kopa tiek veidota atbilstoši izvēlētajai validēšanas metodei. Shēmā tiek pieņemts, ka katrā iterācijā pirms modeļa prognozēšanas kļūdas novērtēšanas tiek veikta labākā modeļa meklēšana. Izmantojot doto shēmu modeļu novērtēšanas procesa attīlošanai modeļu izvēlē, „labākā modeļa meklēšana” tiek aizstāta ar „modeļa parametru aprēķināšana” un testa kopas vietā tiek lietota validēšanas kopa.

Otrā nodaļa. Polinomu regresijā modeļa ar vislabāko prognozēšanas spēju veidošana bieži tiek saukta arī par modeļu būvēšanu [Had2001, Lar2006, Mye2002]. Modeļu būvēšanā

parasti tiek izmantota apakškopas atlases (*subset selection*) pieeja [Had2001, Has2001, Mil2002]. Tās mērķis ir atrast tādu iepriekš definētas pilnās bāzes funkciju kopas apakškopu, kas atbilst regresijas modelim ar vislabāko prognozēšanas spēju dotajos datos. Lai atrastu šādu vislabāko modeli, apakškopas atlases pieejā tiek veikta pārmeklēšana, meklējot modeli, kurš, atbilstoši lietotajam modeļu prognozēšanas spējas novērtējuma mēram, ir vislabākais. Šajā nodaļā tiek apskatīta apakškopas atlase kā stāvokļu telpas pārmeklēšanas problēma un izanalizēti iespējamie tās risinājumi, problēmas un alternatīvas.



4. att. Trīs reizes atkārtots validēšana process: (a) labākā modeļa meklēšana, izmantojot apmācības kopas datus; (b) labākā atrastā modeļa prognozēšanas kļūdas novērtēšana, izmantojot testa kopas datus; (c) labāko atrasto modeļu prognozēšanas kļūdu vidējās vērtības aprēķināšana (shēma adaptēta un vispārināta no [Koh1997])

Ņemot vērā [Aha1994, Joh1994, Mol2002, Web2002], polinomu regresijā apakškopas atlases problēmu var definēt šādi: galīgā fiksētā „pilnā” bāzes funkciju kopā f atrast tādu apakškopu $f^* \subseteq f$, kuras veidotais regresijas modelis

$$F(x) = \sum_{i=1}^k a_i f_i^*(x) \tag{3}$$

starp visiem pārējiem no kopas f potenciāli izveidojamajiem modeļiem ir ar minimālu prognozēšanas kļūdu. Pie tam šāda „labākā” apakškopa var nebūt obligāti unikāla.

Tātad vispirms, pirms labākās apakškopas meklēšanas, ir jānodefinē pilnā kopa f . Tā kā praksē modeļa minimālas prognozēšanas kļūdas sasniegšanai tajā potenciāli iekļaujamo bāzes funkciju kopa parasti nav zināma, tad parasti šī problēma tiek risināta, izvēloties pietiekami augstas kārtas p pilnu polinomu, kurā esošās bāzes funkcijas tad arī kalpos kā kopas f elementi. Šajā gadījumā kopējais visu definēto bāzes funkciju skaits ir

$$m = \prod_{i=1}^p (1 + d/i). \quad (4)$$

Jebkurā f kopas apakškopā $f' \subseteq f$ iekļauto bāzes funkciju sarakstu var uzdot, ar m bināru vērtību vektoru τ , kurā j -tās vērtības vienādība ar 1 nozīmē, ka bāzes funkcija f_j ir iekļauta apakškopā, bet vienādība ar 0 nozīmē, ka nav iekļauta:

$$f' = \{f_j \mid j = 1, 2, \dots, m; \tau_j = 1\}. \quad (5)$$

Formāli labākās apakškopas f^* atlasē problēmu var apskatīt kā m bitu labākās vērtību kombinācijas τ^* izvēles problēmu:

$$\tau^* = \underset{\tau}{\operatorname{arg\,min}} J(\{f_j \mid j = 1, 2, \dots, m; \tau_j = 1\}), \quad (6)$$

kur $J(\cdot)$ ir kritērijs, kas novērtē no apakškopas izveidojamā regresijas modeļa prognozēšanas kļūdu.

Bāzes funkciju apakškopas atlasē problēmu var apskatīt arī kā faktoru atlasē (*feature selection*) jeb faktoru apakškopas atlasē (*feature subset selection*) problēmu, kuras risināšanai pēdējos gados ir pievērsta liela uzmanība (piemēram, [Blu1997, Das1997, Guy2003, Guy2006, Jai1997, Joh1994, Koh1997, Lan1994, Mol2002]). Faktoru apakškopas atlasē optimālās vektora τ bitu vērtību kombinācijas (jeb optimālās faktoru apakškopas) meklēšanu ir ērti apskatīt kā stāvokļu telpas pārmeklēšanu, kur katrs stāvoklis atbilst vienai konkrētai unikālai τ bitu vērtību kombinācijai (jeb vienai konkrētai modelī iekļauto faktoru kombinācijai) [Blu1997, Koh1997, Lan1994].

Lai atrastu apakškopu, kas veido modeli ar vislabāko prognozēšanas spēju, ir nepieciešams veikt kaut kāda veida pārmeklēšanu. Pārmeklēšanas veids, kas vienmēr garantē optimālā risinājuma atrašanu, ir pilnā pārmeklēšana (*exhaustive search*), kas izskata visus iespējamus risinājumus un izvēlas labāko. Tomēr, tā kā, palielinoties m , visu iespējamo apakškopu skaits palielinās eksponenciāli (visu apakškopu skaits, kuras ir iespējams izveidot no m faktoriem, ieskaitot tukšu kopu, ir 2^m), tad arī nepieciešamie skaitļošanas resursi palielinās eksponenciāli. Tādēļ, pat ņemot vērā mūsdienās pieejamās skaitļošanas jaudas, pilnās pārmeklēšanas izmantošana vispārīgā gadījumā ir nepraktiska.

Efektīvāks apakškopas atlasē problēmas risināšanas veids ir izmantot heuristisku pārmeklēšanu. Heuristiskas metodes ir sakņotas pieredzē, racionālās idejās un minējumos. Tās parasti ļauj sasniegt vēlamos rezultātus, bet negarantē to optimalitāti [Rus2002]. Heuristisku metodi var definēt kā metodi, kura meklē labus (t.i., tuvus optimālam) risinājumus, izmantojot pieņemama apjoma skaitļošanas resursus, taču negarantē risinājumu optimalitāti [Ray1996]. Tā vietā, lai izmēģinātu visu iespējamus stāvokļu telpas stāvokļus, galvenais to darbības princips ir, izmantojot noteiktu novērtējuma mēru, censties koncentrēties tikai uz pašiem daudzsoļākajiem stāvokļiem, tādā veidā ievērojami ietaupot skaitļošanas resursus.

Apkopojot [Blu1997, Das1997, Gin1993, Koh1995, Lan1994, Mol2002, Rus2002], var teikt, ka, apskatot apakškopas atlasē kā stāvokļu telpas pārmeklēšanas problēmu, to raksturo šādas piecas pamatīpašības:

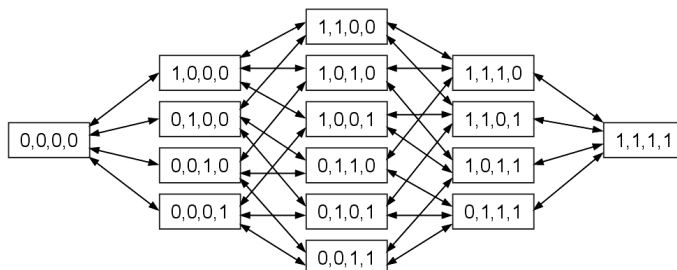
- 1) pārmeklēšanas sākuma stāvoklis;
- 2) stāvokļu pārejas operatori;
- 3) pārmeklēšanas stratēģija;
- 4) stāvokļu novērtējuma mērs;

5) pārmeklēšanas beigu nosacījums.

Promocijas darbā katrā no šīm piecām pamatīpašībām tiek apskatīta tuvāk, izanalizējot un sistematizējot to iespējamās variācijas apakškopas atlasē.

Analizējot apakškopas atlasē visbiežāk lietotos stāvokļu pārejas operatorus, tika konstatēts, ka apakškopas atlasē pamatā tiek lietoti šādi divi pamatoperatori [Mol2002, Koh1997]: pievienošanas operators (viena pašreizējā apakškopā neesoša faktora pievienošana tai) un dzēšanas operators (viena pašreizējā apakškopā esoša faktora dzēšana no tās). Bez pamatoperatoriem atsevišķi vēl tiek nodalīti saliktie un pārleķšanas operatori.

Pārmeklēšanas sākuma stāvoklis kopā ar lietotajiem stāvokļu pārejas operatoriem formē stāvokļu telpu ar visiem stāvokļiem, kuri ir sasniedzami, uzsākot pārmeklēšanu no sākuma stāvokļa un iteratīvi pielietojot operatorus jebkurā secībā [Rus2002]. Stāvokļu telpa ir grafs, kura virsotnes ir stāvokļi, bet loki ir stāvokļu pārejas. Pie tam katrs atsevišķais stāvoklis atbilst vienai konkrētai vektora τ bitu vērtību kombinācijai (t.i., vienai faktoru apakškopai jeb vienam modelim). 5. attēlā ir parādīts stāvokļu telpas piemērs, ja $m = 4$ un tiek lietoti abi pamatoperatori. Lietojot šos abus operatorus, ir iespējams izveidot jebkuru vektora τ bitu vērtību kombināciju no jebkuras citas.



5. att. Apakškopas atlasē stāvokļu telpas piemērs, ja $m = 4$ un tiek lietoti abi pamatoperatori

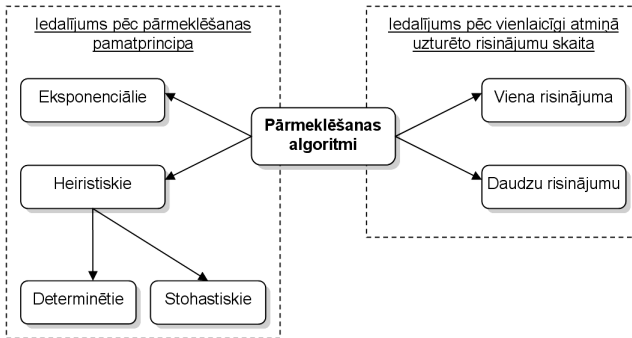
Pārmeklēšanas stratēģija nosaka, kādā veidā notiks pārmeklēšana, izmantojot dotos stāvokļu pārejas operatorus un stāvokļu novērtējuma mēru un ņemot vērā pārmeklēšanas sākuma stāvokli un pārmeklēšanas beigu nosacījumu [Blu1997, Das1997, Mol2002]. Analizējot apakškopas atlasē visbiežāk lietotās pārmeklēšanas stratēģijas, tika konstatēts, ka, ņemot vērā to pārmeklēšanas pamatprincipus, tās var iedalīt trīs kategorijās (skat., piemēram, [Jai1997, Reu2006, Zon1996]):

- 1) eksponenciāla pārmeklēšana;
- 2) determinēta pārmeklēšana;
- 3) stohastiska pārmeklēšana.

Algoritmi, kas ietilpst pirmajā kategorijā, vienmēr garantē rezultātu optimalitāti. Visu eksponenciālo algoritmu pārmeklēšanas laiks ir eksponenciāls, t.i., $O(2^m)$, kas ļauj praktiski pielietot šos algoritmus tikai visvienkāršākajās problēmās. Determinēto algoritmu pārmeklēšanas laiks parasti ir $O(m^2)$ vai pat tikai $O(m)$ [Das1997, Guy2006, Mol2002], taču tie ir īpaši jūtīgi pret lokālajiem minimumiem. Stohastisko algoritmu potenciāli apmeklēto stāvokļu skaits ir tāds pats kā eksponenciālajiem algoritmiem, t.i., $O(2^m)$

[Das1997], taču parasti tiem tiek uzstādīts relatīvi neliels maksimālais izpildāmo iterāciju skaits, neļaujot apmeklēt eksponenciāli lielu stāvokļu skaitu.

6. attēlā ir parādīts pārmeklēšanas algoritmu iedalījums gan pēc to pārmeklēšanas pamatprincipa, gan pēc to vienlaicīgi atmiņā uzturēto risinājumu skaita.



6. att. Pārmeklēšanas algoritmu iedalījums (adaptēts un apvienots no [Jai1997, Mol2002, Zon1996])

Nodaļā ir tuvāk apskatīti apakškopas atlasē populārākie determinētie un stohastiskie pārmeklēšanas algoritmi, tajā skaitā dažādi „kalnā kāpšanas” (*hill climbing*) algoritmu paveidi, peldoša pārmeklēšana (*floating search*), starveida pārmeklēšana (*beam search*), simulētā atkvēlināšana (*simulated annealing*), ģenētiskie algoritmi u.c. Secināts, ka modeļu būvēšanas praksē vispopulārākais algoritms ir „Secīga uz priekšu vērsta atlase” (*Sequential Forward Selection, SFS*) [Aha1996, Kit1978], saukts arī par „Uz priekšu vērstu atlasī” (*Forward Selection*) vai par „Uz priekšu vērstu soļveida atlasī” (*Forward Stepwise Selection*) [Has2001, Lar2006, Mil2002].

Tālāk nodaļā tiek apskatīta apakškopas atlase ortogonālo polinomu regresijā, cita veida stāvokļu pārejas operatoru izmantošana, pārmeklēšanas intensitātes izraisīta pārmērīga piemērošanās, kā arī modeļu kombinēšana.

Trešajā nodaļā tiek apskatīti apakškopas atlases pieejas trūkumi, kas ir saistīti ar pilnās bāzes funkciju kopas f veidošanu un tās labākās apakškopas meklēšanai nepieciešamajiem skaitļošanas resursiem praktiskos pielietojumos.

Apakškopas atlases pieejā tiek meklēta pilnās bāzes funkciju kopas apakškopa, kurai atbilstošais modelis ir ar maksimālu prognozēšanas spēju. Tas nozīmē, ka apakškopas atlases pieejā tiek pieņemts, ka šajā pilnās bāzes funkciju kopas veidošanas procesā izveidotā pilnā fiksētā kopa satur apakškopu, kas ir pietiekama, lai pietiekami labi aprakstītu pētāmās sakarības.

Dažādām ar modeļiem aprakstāmajām sakarībām var būt atšķirīga uzvedība un atšķirīga rezultatīvās pazīmes y nelinearitāte, t.i., to aprakstīšanai var būt nepieciešamas atšķirīgu formu un pakāpju bāzes funkciju kombinācijas, tādēļ pilnajai bāzes funkciju kopai ir jābūt vai nu vienmēr pietiekami lielai, lai tajā atrastos visas potenciāli nepieciešamās bāzes funkcijas (praktiskos pielietojumos vispārīgā gadījumā neizpildāma prasība), vai arī tā katru

reizi ir speciāli jāpiemeklē risināmajai problēmai, izmēģinot dažādus variantus un katru reizi atkārtojot apakškopas atlasē pārmeklēšanas procesu no paša sākuma.

Tā kā praktiskos pielietojumos modeļa prognozēšanas kļūdas minimuma sasniegšanai potenciāli nepieciešamo konkrēto bāzes funkciju apakškopa parasti nav zināma (īpaši maz izpētītās problēmsfērās ar ļoti ierobežotu zināmo modelēšanai pielietojamo informāciju), tad pilnās kopas izvēles problēma tiek risināta vienkāršāk – izvēloties pietiekami augstas kārtas p pilnu polinomu, kurā esošās m bāzes funkcijas tad arī kalpos kā pilnās kopas elementi. Tādā gadījumā katrai risināmajai modelēšanas problēmai pirms apakškopas atlasē ir jāpiemeklē tikai polinoma maksimālā kārtā p .

Lai gan kārtas piemeklēšanas problēma ir vienkāršāka par atsevišķu bāzes funkciju piemeklēšanu, tomēr problēmas netrivialitāte saglabājas – dažādām sakarībām var būt nepieciešamas atšķirīgas p vērtības. Tradicionāli, šīs problēmas risinājums ir bijis pavisam pragmatisks – balstīts uz intuīciju, minēšanu, iepriekšējo pieredzi vai arī vienkārši balstīts uz pieejamās statistiskās programmatūras dotajām iespējām.

Jebkurā gadījumā šāda pilnās kopas vai kārtas piemeklēšana daudzos gadījumos var izvērsties sarežģītā un ilgā mēģinājumu un kļūdu procesā, jo ar apakškopas atlasē metodēm iegūto modeļu prognozēšanas spēja var būt ļoti jutīga pret pilnās bāzes funkciju kopas izvēli [Rik1999, Rik2003, Tod2003, Tod2004, Jek2008].

Bez tam p vērtības izvēli, praktisku apsvērumu dēļ, ietekmē arī modeļa būvēšanas procesam nepieciešamo skaitļošanas resursu apjoms. Pielietojot vienu vai abus stāvokļu pārejas pamatoperatorus, polinomu regresijas apakškopas atlasē stāvokļu telpas atsevišķa stāvokļa zarošanās koeficients ir vienāds ar kopējo bāzes funkciju skaitu, kas savukārt, atkarībā no oriģinālo faktoru skaita d un maksimālās kārtas p , palielinās ar kārtu $O(m) = O(d^p)$ [Has2001, Lin2007, Tod2003, Tod2004]. Pie tam tas ir tik liels jau pašā pirmajā pārmeklēšanas iterācijā. Tas nozīmē, ka, palielinoties y nelinearitātei un faktoru skaitam, modeļa, kas apraksta pētāmās sakarības pietiekami labi (tātad modeļa ar augstu kārtu), atrašana var izrādīties praktiskiem pielietojumiem pārmērīgi laikietilpīgs process, pat, izmantojot visvienkāršākos determinētos heuristiskās pārmeklēšanas algoritmus. Tas savukārt nozīmē, ka, praktiskos pielietojumos rezultātu sasniegšanai pieņemamā laikā, palielinoties d , ir nepieciešams samazināt p , lai gan nav pamata uzskatīt, ka vispārīgā gadījumā y nelinearitāte patiešām samazināsies, palielinoties d .

Apskatīto trūkumu novēršanai ir nepieciešama no apakškopas atlasē pieejas atšķirīga pieeja, kura neprasītu iepriekš sagatavotu fiksētu pilno bāzes funkciju kopu (vai modeļu maksimālās kārtas uzstādīšanu), bet gan nodrošinātu automātisku adaptīvu jebkuru pietiekami precīzai datu aprakstīšanai nepieciešamo bāzes funkciju konstruēšanu pārmeklēšanas laikā un tādā veidā ļautu efektīvi ģenerēt regresijas modeļus ar augstu prognozēšanas spēju neatkarīgi no konkrēto pētāmo sakarību veiksmīgai aprakstīšanai nepieciešamo bāzes funkciju formas.

Ceturtajā nodaļā tiek piedāvāta adaptīvas bāzes funkciju konstruēšanas (*Adaptive Basis Function Construction*, ABFC) pieeja. ABFC pieeju var uzskatīt par alternatīvu apakškopas atlasē pieejai regresijas modeļu būvēšanā.

Pretstatā apakškopas atlasē pieejai, kurā modeļu būvēšana notiek, meklējot labāko bāzes funkciju kombināciju no galīga izmēra fiksētas iepriekš izveidotas pilnās bāzes funkciju kopas, ABFC pieejā modeļa būvēšanai nepieciešamās konkrētās bāzes funkcijas nav

iepriekš jādefinē – tās tiek adaptīvi ģenerētas heuristiskas pārmeklēšanas gaitā, balstoties uz konkrētās risināmās modelēšanas problēmas datiem.

Polinomu regresijas modeli var definēt kā bāzes funkciju summu (kā vienādojumā (2)):

$$F(x) = \sum_{i=1}^k a_i f_i(x), \quad (7)$$

kur $f_i(x)$, $i = 1, 2, \dots, k$ ir polinoma bāzes funkcijas, kuras vispārīgā veidā var tikt definētas kā noteiktā pakāpē kāpinātu faktoru reizinājums:

$$f_i(x) = \prod_{j=1}^d x_j^{r_{i,j}}, \quad (8)$$

kur r ir $k \times d$ izmēru faktoru pakāpju matrica: $r_{i,j}$ ir i -tās bāzes funkcijas j -tā faktora pakāpe (nenegatīvs vesels skaitlis). Ja noteiktai i -tajai bāzes funkcijai $\forall j: r_{i,j} = 0$, tad šī bāzes funkcija atbilst polinoma brīvajai konstantei.

Atšķirībā no apakškopas atlasēšanas problēmas, kur vektora τ katrs bits parāda tikai, vai tam atbilstošā bāzes funkcija ir iekļauta modelī, bet pašas bāzes funkcijas ir definētas atsevišķi, ABFC pieejā, pie dota faktoru skaita d , matrica r , ar konkrētu rindu skaitu k un konkrētām tās elementu vērtībām, pilnībā definē modeli. Modelī iekļauto bāzes funkciju kopa f ir vienāda ar

$$f = \left\{ \prod_{j=1}^d x_j^{r_{i,j}} \mid i = 1, 2, \dots, k \right\}. \quad (9)$$

Piemēram, ja $d = 3$ un $k = 4$, tad matrica

$$r = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix}$$

atbilst bāzes funkciju kopai

$$f = \left\{ x_1^0 x_2^0 x_3^0, x_1^1 x_2^0 x_3^0, x_1^0 x_2^1 x_3^3, x_1^1 x_2^1 x_3^1 \right\} = \left\{ 1, x_1, x_2 x_3^3, x_1 x_2 x_3 \right\}$$

jeb modelim

$$\begin{aligned} F(x) &= a_1 x_1^0 x_2^0 x_3^0 + a_2 x_1^1 x_2^0 x_3^0 + a_3 x_1^0 x_2^1 x_3^3 + a_4 x_1^1 x_2^1 x_3^1 = \\ &= a_1 + a_2 x_1 + a_3 x_2 x_3^3 + a_4 x_1 x_2 x_3. \end{aligned}$$

Formāli ABFC pieejā labākās modelī iekļaujāmās bāzes funkciju kopas f^* izvēles problēmu var apskatīt kā matricas r elementu vērtību (un tās rindu skaita) labākās kombinācijas r^* izvēles problēmu:

$$r^* = \arg \min_r J \left(\left\{ \prod_{j=1}^d x_j^{r_{i,j}} \mid i = 1, 2, \dots, k \right\} \right), \quad (10)$$

kur $J(\cdot)$ ir kritērijs, kas novērtē no bāzes funkciju kopas izveidojamā regresijas modeļa prognozēšanas kļūdu; k ir matricas r rindu skaits jeb modelī iekļauto bāzes funkciju skaits, kas tiek piemeklēts kopā ar matricas elementu vērtībām. Tā kā ne matricas elementu vērtību,

ne rindu skaita k augšējās robežas netiek definētas, tad tiek iegūta bezgalīga modeļu kandidātu telpa un ir iespējams ģenerēt jebkuras sarežģītības polinomus bez konkrētu bāzes funkciju iepriekšējas definēšanas.

Lai pārmeklēšanas procesā vienmēr būtu iespējams izveidot jebkuru bāzes funkciju kopu, kuru atļauj matricas r definīcija, ir nepieciešami atbilstoši stāvokļu pārejas operatori. ABFC pieejā tiek izmantoti tādi operatori, kas ļauj ne tikai pievienot modelim jaunas bāzes funkcijas un dzēst tās (kā tas ir apakškopas atlasēs pieejā, radot nepieciešamību definēt iespējamās bāzes funkcijas), bet ļauj arī modificēt modeli jau esošās bāzes funkcijas, kā arī veidot to modificētas kopijas. Tādā veidā visas nepieciešamās bāzes funkcijas tiek adaptīvi konstruētas pārmeklēšanas procesa gaitā.

ABFC pieejā ir iespējams izstrādāt daudz dažādus stāvokļu pārejas operatorus. Tomēr ir svarīgi, lai ar tiem izdarītās modeļu modifikācijas būtu pietiekami nelielas, neļaujot stāvokļu telpā ar lielu lēcieni pārlēkt pāri potenciāli labākajiem risinājumiem vai tādiem risinājumiem, kas potenciāli var novest pie labākajiem risinājumiem, kā arī, neļaujot pārmērīgi palielināt stāvokļu telpas zarošanās koeficientu, kas izraisītu pārmeklēšanai nepieciešamo skaitļošanas resursu strauju pieaugumu. Tiek piedāvāta piecu operatoru kopa, kuru lietojot ir iespējams izveidot visus iespējamus nenegatīvas kārtas polinomu modeļus neatkarīgi no izvēlētā pārmeklēšanas sākuma stāvokļa:

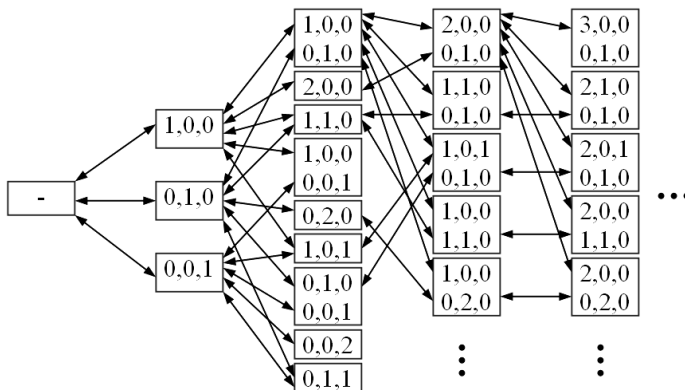
- Operators1: Vienas jaunas bāzes funkcijas pievienošana ar vienu no pakāpēm uzstādītu vienādu ar 1 (pārējās – vienādas ar 0).
- Operators2: Vienai modelī esošai bāzes funkcijai vienas no pakāpēm palielināšana par 1.
- Operators3: Vienas modelī esošās bāzes funkcijas identiskas kopijas, kurai viena no pakāpēm ir palielināta par 1, pievienošana modelim.
- Operators4: Vienai modelī esošai bāzes funkcijai vienas no pakāpēm, kura ir lielāka par 0, samazināšana par 1.
- Operators5: Vienas modelī esošās bāzes funkcijas dzēšana.

Pārmeklēšanas sākuma stāvoklis kopā ar lietotajiem stāvokļu pārejas operatoriem formē stāvokļu telpu. 7. attēlā ir parādīts neliels ABFC stāvokļu telpas sākuma daļas piemērs, manipulējot ar matricu r , ja $d = 3$ un pārmeklēšana tiek uzsākta no visvienkāršākā modeļa. Šādā telpā sarežģītības operatori ģenerē stāvokļu pārejas sarežģītāku modeļu virzienā (attēlā – pa labi), bet vienkāršošanas operatori – vienkāršāku modeļu virzienā (attēlā – pa kreisi).

Uzsākot pārmeklēšanu no stāvokļu telpas vienkāršākā modeļa, apakškopas atlasēs pieejā, lai arī modeļu sarežģītība stāvokļu telpā aug atkarībā no telpas slāņa (katrā nākamajā slānī modeļos ir par vienu bāzes funkciju vairāk), tomēr jau pašā pirmajā iterācijā ir iespējams izvēlēties modeli ar iekļautu kādu no vissarežģītākajām iespējamajām bāzes funkcijām (t.i., ar vislielākajām summārajām faktoru pakāpēm). ABFC pieejā turpretī, stāvokļu telpa ir sakārtota tādā veidā, ka telpas sākuma daļā pirmkārt ir atrodami gan tikai visvienkāršākie modeļi, gan arī tikai visvienkāršākās bāzes funkcijas. Tas dod iespēju pirmkārt koncentrēties ne tikai uz modeļiem ar mazāku iekļauto bāzes funkciju skaitu, bet arī uz vienkāršākām bāzes funkcijām, vienlaicīgi ietaupot skaitļošanas resursus.

Pielietojot vienu vai abus stāvokļu pārejas pamatoperatorus, polinomu regresijas apakškopas atlasēs stāvokļu telpas atsevišķa stāvokļa zarošanās koeficients ir vienāds ar kopējo bāzes funkciju skaitu, kas savukārt atkarībā no faktoru skaita d un uzstādītās maksimālās kārtas p palielinās ar kārtu $O(m) = O(d^p)$. Pie tam tik liels tas ir jau pašā pirmajā pārmeklēšanas iterācijā. ABFC pieejā, pielietojot visus piecus piedāvātos stāvokļu

pārejas operatorus, stāvokļu zarošanās koeficienta augšējā robeža ir $O(dk)$, kas ir lineāra gan pret d , gan pret k .



7. att. Neliels ABFC stāvokļu telpas sākuma daļas piemērs. Katrs stāvoklis atbilst vienam konkrētam modelim. Vienkāršības labad, šeit nav parādīta matricas r pirmā rinda, kura vienmēr atbilst brīvajai konstantei (rindā visas nulles), kā arī nav parādītas starpsplāņu stāvokļu pārejas, kuras veido Operators3

Pieņemot, ka pārmeklēšanas rezultātā atrastā modeļa bāzes funkciju skaits ir k_* un katrā pārmeklēšanas iterācijā modeļu sarežģītība palielinās par 1, izmantojot apakškopas atlasas pieeju ar vienu vai abiem tās pamatoperatoriem, izskatītais stāvokļu skaits būs vienāds ar $O(d^p k_*)$, taču, izmantojot ABFC pieeju, izskatītais stāvokļu skaits būs vienāds ar $O(dk_*^3)$. Tātad ir sagaidāms, ka modeļu būvēšanai nepieciešamo skaitļošanas resursu kontekstā, relatīvi pret apakškopas atlasas pieejas efektivitāti, ABFC pieejas efektivitāte pieaugs līdz ar faktoru skaita un uzstādītās modeļu maksimālās kārtas palielināšanos un meklējamā modeļa sarežģītības k_* samazināšanos.

Ir svarīgi piebilst, ka, lai gan ABFC pieejā stāvokļu telpa ir bezgalīgi liela, bet tās pārmeklēšanai tiek veltīts tikai pavisam neliels galīgs laiks, tas nenozīmē, ka modeļu ar pietiekami augstu prognozēšanas spēju atrašanai vispārīgā gadījumā ir bezgalīgi maza varbūtība. Tā iemesls ir tas, ka tieši visvienkāršākie modeļi ar vismazāko skaitu un visvienkāršākajām bāzes funkcijām – tie, kas atrodas vistuvāk pieņemtajam sākuma stāvoklim, ar vislielāko varbūtību arī ir tie, kuru prognozēšanas spēja ir sagaidāma vislabākā.

Apakškopas atlasas pieejas un ABFC pieejas raksturīgākās atšķirīgās īpašības ir apkopotas 1. tabulā.

1. tabula

Apakškopas atlasas pieejas un ABFC pieejas raksturīgāko īpašību salīdzinājums

Īpašība	Apakškopas atlasas pieeja	ABFC pieeja	Piezīmes par ABFC pieeju
Modeļu būvēšanai paredzētās bāzes funkcijas	Iepriekš jāizveido pilnā kopa vai jāpiemeklē maksimālā kārtā p (iespējams, vairākkārtīgi atkārtojot pārmeklēšanas procesu)	Bāzes funkcijas tiek adaptīvi konstruētas pārmeklēšanas gaitā	Nav jāizvēlas ne pilnā bāzes funkciju kopa, ne maksimālā kārtā (attieciņi nav arī vairākkārtīgi jāatkārto pārmeklēšanas process)
Modeļu būvēšanas pamatprincips	Tiek meklēta labākā pilnās bāzes funkciju kopas apakškopa	Bāzes funkcijas tiek adaptīvi konstruētas pārmeklēšanas gaitā	Vajadzīgās bāzes funkcijas tiek iegūtas kā daļa no risinājuma; tiek iegūta augstāka elastība
Modeļu kārtas ierobežošana un/vai konkrētu iespējamo bāzes funkciju definēš.	Obligāta nepieciešamība	Neobligāta papildu iespēja	
Iespējamais stāvokļu (modeļu) attēlojums	Bināru vērtību vektors ar garumu m	Nenegatīvu veselu skaitļu $k \times d$ matrica	
Pārmeklēšanas sākuma stāvoklis	Jebkurš telpas stāvoklis (tukša kopa, pilna kopa, nejauša kopa)	Telpas visvienkāršākais modelis vai kāds cits pietiekami vienkāršs modelis	Uzstādot attiecīgus ierobežojumus, ir iespējams uzsākt arī no jebkura cita stāvokļa
Stāvokļu pārejas operatori	Pievienošanas, dzēšanas, saliktie, pārlekšanas	Sarežģīšanas, vienkāršošanas, saliktie, pārlekšanas	Ir iespējams izstrādāt arī dažāda veida citus operatorus
Pārmeklēšanas stratēģija	Visi promocijas darbā apskatītie (tipiskie) algoritmi	Visi apskaitītie algoritmi izņemot tos, kuri prasa konstantu stāvokļu attēlojuma izmēru	
Stāvokļu novērtējuma mērs	Statistiskā hipotēžu testēšana, sarežģītības sodīšanas kritēriji, validēšanas metodes	Sarežģītības sodīšanas kritēriji un validēšanas metodes	
Beigu nosacījums	Visi promocijas darbā uzskaitītie	Visi uzskaitītie, izņemot „apstāšanās telpas pretējā pusē”	Ar attiecīgiem ierobežojumiem ir iespējams izmantot arī šo nosacījumu
Stāvokļu skaits telpā (jeb iespējamo bāzes funkciju kombināciju skaits)	$O(2^m)$ jeb $O(2^{d^p})$	Bezgalīgs skaits	Ja nepieciešams, uzstādot attiecīgus ierobežojumus, arī ABFC pieeja tiek iegūts to galīgs skaits
Stāvokļu telpas stāvokļa zarošanās koeficients	$O(d^p)$ – eksponenciāli augošs	$O(dk)$ – lineāri augošs	

Izskatīto stāvokļu skaits, ja sameklētā modeļa bāzes funkciju skaits ir k_*	$O(d^p k_*)$ – eksponenciāli augošs	$O(dk_*^3)$ – polinomiski augošs	
-------------------------------------------------------------------------------	-------------------------------------	----------------------------------	--

Izstrādātā vispārīgā ABFC pieeja var kalpot kā universāls karkass, uz kura pamata ir iespējams izstrādāt dažādas jaunas adaptīvas regresijas modeļu būvēšanas metodes, ar šādām (no apakškopas atlasēs metodēm atšķirīgām) īpašībām:

- nav nepieciešams definēt pilno bāzes funkciju kopu vai izvēlēties modeļu maksimālo kārtu;
- ir iespējams ģenerēt neierobežotas sarežģītības un kārtas modeļus;
- modeļu būvēšanai nepieciešamo bāzes funkciju konstruēšana, un līdz ar to arī sarežģītība un kārtā, tiek noteikta pārmeklēšanas gaitā;
- izveidojamo modeļu iespējamo funkcionālo formu paaugstinātās elastības dēļ, ir iegūta paaugstināta spēja piemēroties datos esošām nelineārām daudzfaktoru sakarībām;
- salīdzinājumā ar apakškopas atlasēs metodēm, ir iegūtas samazinātas skaitļošanas resursu prasības.

Promocijas darbā tiek piedāvāti arī divi ABFC pieejas speciāli gadījumi – divas konkrētas regresijas modeļu būvēšanas metodes:

- Peldoša adaptīva bāzes funkciju konstruēšana (*Floating Adaptive Basis Function Construction*, F-ABFC) – realizē adaptīvu bāzes funkciju konstruēšanu, izmantojot peldošas uz priekšu vērstas pārmeklēšanas principus [Pud1994a, Pud1994b] un koriģēto Akaiķes informācijas kritēriju [Aka1973, Aka1974] (2. tabulā ir parādīts F-ABFC pseidokods);
- Peldošas adaptīvas bāzes funkciju konstruēšanas ansamblis (*Ensemble of Floating Adaptive Basis Function Construction*, EF-ABFC) – metode tika izstrādāta kā F-ABFC paplašinājums (skat. 8. attēlu), lai cīnītos pret problēmām, kas saistītas ar intensīvas pārmeklēšanas izraisītu pārmērīgu piemērošanos un modeļu izvēles dispersiju, izmantojot šķērsvalidēšanu un modeļu vidējošanu. Salīdzinot ar F-ABFC metodi, tās trūkums ir palielinātie modeļu būvēšanai nepieciešamie skaitļošanas resursi. Taču to ir iespējams mazināt, jo EF-ABFC ļauj ļoti vienkāršā veidā veikt skaitļošanas paralelizēšanu.

Izmantojot izstrādātās ABFC metodes, pirms modeļu būvēšanas procesa uzsākšanas lietotājam nav nepieciešams uzstādīt nekādu hiperparametru vērtības, tādā veidā ievērojami vienkāršojot modelēšanas procesu. Hiperparametru uzstādīšana šīm metodēm ir pārtapusi par neobligātu papildu iespēju – piemēram, vajadzības gadījumā, ja ir pieejama nepieciešamā informācija, ir iespējams ierobežot atsevišķu bāzes funkciju vai visu modeļu pakāpes līdzīgā veidā, kā tas notiek apakškopas atlasēs pieejā.

ABFC pieeja ir pielietojama ne tikai polinomu regresijā ar nenegatīvu pakāpju bāzes funkcijām, bet tajā skaitā arī ar cita veida bāzes funkcijām, polinomu neironu tīklos, ar nelineārās regresijas modelēšanas metodēm, kā arī klasificēšanas problēmu risināšanā.

Piektā nodaļa ir veltīta empīriskiem eksperimentiem, kuros tiek novērtēta promocijas darbā izstrādāto ABFC metožu efektivitāte praktiskos pielietojumos, salīdzinot tās ar citām plaši lietotām regresijas modelēšanas metodēm. Šajā nodaļā ABFC pieejas metodes tiek

pielietotas regresijas modeļu būvēšanai regresijas datu kopām no dažādām pielietojumu sfērām, salīdzinot tās ar pilno polinomu regresiju un polinomu regresijas standarta apakškopas atlasē metodēm. Salīdzinot iegūtos rezultātus, galvenā uzmanība tiks pievērsta iegūto modeļu prognozēšanas spējai, kā arī ABFC pieejas metožu ātrdarbības ieguvumam. Papildus, lai noteiktu ABFC pieejas metožu konkurētspēju, tiek veikti salīdzinājumi arī ar plaši zināmām un lietotām regresijas modelēšanas metodēm, kas nesakņojas lineārajā polinomu regresijā.

2. tabula

F-ABFC pārmeķlēšanas procedūras pseidokods

LabākaisModelis := modelis ar brīvajai konstantei atbilstošu bāzes funkciju ($r_{1,j} := 0, j = 1...d$)

LabākaisModelis.AICC := Novērtēt(*LabākaisModelis*)

repeat

 //uz priekšu vērstā fāze

JAUNIEMODEĻI := {visi modeļi, kurus ir iespējams izveidot, izmantojot *LabākaisModelis* un sarežģītāšanas operatorus}

PašreizējaisLabākaisModelis := *LabākaisModelis*

foreach (*TestaModelis* ∈ *JAUNIEMODEĻI*) **do**

TestaModelis.AICC := Novērtēt(*TestaModelis*)

if (*TestaModelis*.AICC < *PašreizējaisLabākaisModelis*.AICC) **then**

PašreizējaisLabākaisModelis := *TestaModelis*

endfor

LabākaisModelis := *PašreizējaisLabākaisModelis*

if (*LabākaisModelis* nav mainījies) **then**

exit

 //atpakaļ vērstā fāze

repeat

JAUNIEMODEĻI := {visi modeļi, kurus ir iespējams izveidot, izmantojot *LabākaisModelis* un vienkāršošanas operatorus}

foreach (*TestaModelis* ∈ *JAUNIEMODEĻI*) **do**

TestaModelis.AICC := Novērtēt(*TestaModelis*)

if (*TestaModelis*.AICC < *LabākaisModelis*.AICC) **then**

LabākaisModelis := *TestaModelis*

endfor

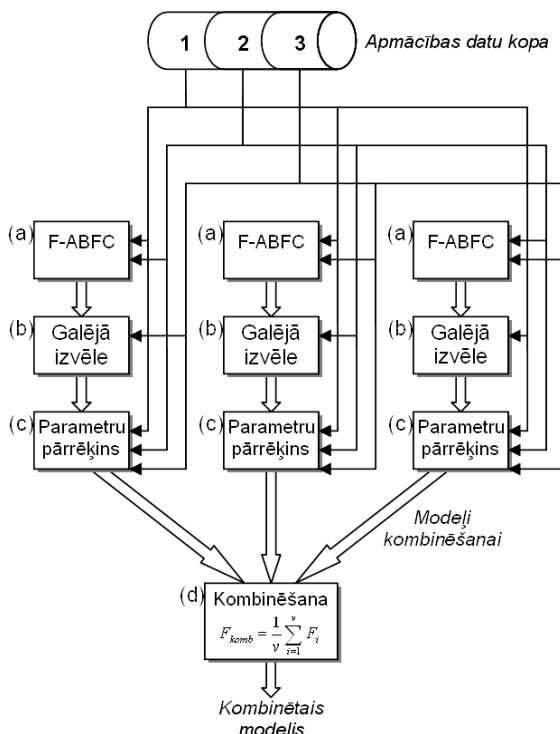
until (*LabākaisModelis* nav mainījies)

until forever

Veiktajos empīriskajos eksperimentos tika salīdzinātas šādas regresijas modelēšanas metodes:

- pilno polinomu regresija;
- polinomu regresijas standarta apakškopas atlasē metode: SFS algoritms un statistiskā hipotēžu testēšana F-testa veidā (SFS + F-tests);
- polinomu regresijas standarta apakškopas atlasē metode: SFS algoritms un AICC modeļu novērtēšanas kritērijs (SFS + AICC);
- M5' regresijas koki (*M5' Regression Trees*, M5' RT) [Qui1992, Wah1997];

- M5' modeļu koki (*M5' Model Trees*, M5' MT) [Qui1992, Wah1997];
- radiālo bāzes funkciju neironu tīkli (*Radial Basis Function Neural Networks*, RBFNN) [Bis1995];
- tuvāko kaimiņu metode (*k-Nearest Neighbour*, k-NN) [Aha1991];
- daudzfaktoru adaptīvie regresijas splaini (*Multivariate Adaptive Regression Splines*, MARS) [Fri1991, Fri1993];
- lokāli svērtie polinomi (*Locally Weighted Polynomials*, LWP) [Clv1995, Fan1996];
- Ierobežota polinomu vienādojumu inducēšana regresijā (*Constrained Induction of Polynomial Equations for Regression*, CIPER) ar minimālā apraksta garuma (*Minimum Description Length*, MDL) kritēriju (CIPER + MDL) [Tod2003, Tod2004];
- CIPER metode ar AICC kritēriju (CIPER + AICC);
- promocijas darbā izstrādātā F-ABFC metode;
- promocijas darbā izstrādātā EF-ABFC metode.



8. att. EF-ABFC modelēšanas procesa kopējā shēma, lietojot trīs daļu šķērsvaidēšanu: (a) labākā modeļa konstruēšana, izmantojot F-ABFC; (b) viena galējā labākā modeļa izvēle, ņemot vērā validēšanas kopas kļūdu; (c) galējā labākā modeļa parametru pārrēķināšana, izmantojot visu apmācības kopu; (d) izveidoto modeļu kombinēšana

Regresijas modelēšanas metožu efektivitātes objektīvai salīdzināšanai tika izmantotas 20 regresijas datu kopas, kurās faktoru skaits d ir robežās no 4 līdz 122 un kopējais piemēru skaits n ir robežās no 73 līdz 950. Šīs kopas ir izvēlētas no dažādām pielietojumu sfērām ar mērķi izskatīt pietiekami plašu faktoru skaita un rezultatīvās pazīmes nelinearitātes un trokšņu līmeņu diapazonu, kā arī dēļ to relatīvi nelielā piemēru skaita, kā tas ir bieži sastopams reālos praktiskos pielietojumos. Pirmās 14 datu kopas ir ņemtas no publiskiem datu kopu repozitorijiem, bet pēdējās 6 kopas pārstāv metamodelēšanas¹ pielietojumu sfēru.

Veiktajos empiriskajos eksperimentos salīdzinātās regresijas modelēšanas metodes tika vērtētas pēc šādiem kritērijiem:

- izveidotā modeļa prognozēšanas kļūda – vidējās kvadrātiskās kļūdas kvadrātsaknes relatīvā vērtība (*Relative Root Mean Squared Error*, RRMSE), kas aprēķināta testa datu kopā;
- modeļa izveidošanai patērētais laiks – tas parāda modelēšanas metodes efektivitāti skaitļošanas resursu kontekstā.

Lai šos novērtējumus aprēķinātu pietiekami ticami, tika izmantota 10 un 5 daļu šķēršvalidēšana. Visos gadījumos validēšana tika veikta „ārējā ciklā” ap katru modelēšanas metodi, katrā no validēšanas iterācijām sadalot kopējos datus apmācības kopā un testa kopā, veicot atkārtotu modeļa būvēšanu, nosakot modeļa veidošanai patērēto laiku un izveidoto modeļu bāzes funkciju skaitu, kā arī aprēķinot RRMSE novērtējumu, izmantojot testa kopas datus, kuri modeļa būvēšanā netika izmantoti.

Tā kā, pielietojot apakškopas atlasē pieeju, lietotāja-eksperta intuīcijas, pieredzes un minējumu virzītu maksimālās kārtas izvēli vispārīgā gadījumā nav iespējams formalizēt, tad maksimālās kārtas izvēles process tika simulēts šādos trīs formalizējamās veidos:

- vienmēr konstanta maksimālā kārtā ($p = 1$, $p = 2$ un $p = 3$);
- maksimālās kārtas izvēle, izmantojot AICC kritēriju – sākot no $p = 1$ un palielinot to, kamēr vien AICC vērtība uzlabojas („SFS + AICC + AICC”);
- maksimālās kārtas izvēle, izmantojot šķēršvalidēšanas procesā iegūto vidējo RRMSE kļūdu – sākot no $p = 1$ un palielinot to, kamēr vien RRMSE vērtība uzlabojas („SFS + AICC + CV”). Sagaidāms, ka tas būs aptuveni v reizes ilgāks process par iepriekšējā punktā aprakstīto, kur v ir šķēršvalidēšanas daļu skaits.

Otrais no šiem trim maksimālās kārtas izvēles veidiem attiecas tikai uz SFS + AICC metodi, jo ar to ir iegūta katrai modeļa kārtai atbilstoša AICC vērtība. Jāpiebilst, ka ar šo veidu iegūtie rezultāti lielākajai daļai apskatīto datu kopu ir nedaudz optimistiski, t.i., tiek uzrādītas mazākas prognozēšanas kļūdas un mazāks laika patēriņš nekā tas būtu patiesībā, jo atsevišķai pārmeklēšanai nepieciešamajiem laika resursiem sasniedzot vairākas stundas, tālāk modeļu maksimālā kārtā vairs netika palielināta pat tādā gadījumā, ja, ņemot vērā izmantoto kritēriju, labākais modelis vēl nebija atrasts (tātad tika ietaupīti skaitļošanas

¹ Metamodelēšanā sarežģīti un skaitļošanas ietilpīgi simulāciju modeļi, izmantojot ar tiem ģenerētas relatīvi neliela apjoma ieejas-izejas datu kopas, tiek aizstāti ar vienkāršākiem un ātrdarbīgākiem (regresijas) modeļiem (sauktiem arī par metamodeļiem vai surogātmodeļiem), ar mērķi tos pielietot pētāmo konstrukciju optimizēšanā, jūtīguma analizē vai „kas notiks, ja...?” (*what-if*) analizē simulāciju modeļu vietā [Chp2006, Mye2002, Rik1999, Rik2003, Wag2007]. Metamodelēšanā visbiežāk lietotais metamodelis ir pilnais kvadrātiskais polinoms, tomēr aizvien vairāk uzmanība tiek pievērsta sarežģītākām regresijas modelēšanas metodēm [Chp2006, Fuj2006, Kal2008a, Sim2001, Wag2007].

resursi un potenciāli samazināta intensīvas pārmeklēšanas izraisītā pārmērīgā piemērošanās).

3. un 4. tabulā ir parādītas salīdzināto regresijas modelēšanas metožu iegūto modeļu vidējās prognozēšanas kļūdas un vidējie modelēšanai patērētie laiki atsevišķi pa pirmajām 14 datu kopām, pa pēdējām 6 datu kopām un kopā pa visām datu kopām. Pilno polinomu regresijas, SFS + F-tests un LWP rezultāti ir norādīti tikai pēdējo 6 kopu kolonnā, jo daļā eksperimentu ar pirmajām 14 datu kopām iegūt modeļus ar šīm metodēm nebija iespējams. Apakškopas atlases (kopā ar maksimālās kārtas piemeklēšanu) metožu rezultāti un promocijas darbā piedāvāto adaptīvās bāzes funkciju konstruēšanas pieejas metožu rezultāti abās tabulās ir izcelti treknrakstā.

3. tabula

Salīdzināto regresijas modelēšanas metožu iegūto modeļu vidējās prognozēšanas kļūdas

Metode	Vidējā RRMSE kļūda (pirmās 14 kopas)	Vidējā RRMSE kļūda (pēdējās 6 kopas)	Vidējā RRMSE kļūda (visas kopas)
Pilnais polin., $p = 1$	-	56.45	-
Pilnais polin., $p = 2$	-	34.04	-
Pilnais polin., $p = 3$	-	274.60	-
Pilnais polin. + CV	-	12.91	-
SFS + F-tests, $p = 1$	-	56.71	-
SFS + F-tests, $p = 2$	-	33.57	-
SFS + F-tests + CV	-	10.80	-
SFS + AICC, $p = 1$	53.79	56.74	54.67
SFS + AICC, $p = 2$	47.39	33.58	43.25
SFS + AICC, $p = 3$	-	23.02	-
SFS + AICC + AICC	57.66	8.75	42.98
SFS + AICC + CV	41.93	8.30	31.84
M5' RT	59.92	66.73	61.96
M5' MT	41.91	22.80	36.18
RBFNN	5E+10 (72.15)	75.59	3E+10 (73.24)
k-NN	48.48	50.92	49.21
MARS	46.52	12.62	36.35
LWP	-	12.23	-
CIPER + MDL	44.33	24.94	38.51
CIPER + AICC	54.30	19.41	43.83
F-ABFC	50.47	6.07	37.15
EF-ABFC	38.30	6.07	28.63

Apskatot pilno polinomu un apakškopas izvēles metožu rezultātus, var secināt, ka vismazākās prognozēšanas kļūdas ir iegūtas, piemeklējot maksimālo kārtu ar šķērsvārdēšanu (SFS + AICC + CV). Taču šāda veida piemeklēšana parasti ir arī vislaikietilpīgākā.

F-ABFC dod iespēju veikt modeļu būvēšanu, izmantojot ievērojami mazākus skaitļošanas resursus – aptuveni 22 reizes mazākus, salīdzinot ar SFS + AICC + AICC, un aptuveni 69 reizes mazākus, salīdzinot ar SFS + AICC + CV. Tomēr, vismaz atsevišķos gadījumos, ar F-ABFC metodi iegūtie modeļi ir pārmērīgi piemērojušies, pasliktinot kopējos rezultātus.

Ar EF-ABFC metodi iegūto modeļu prognozēšanas kļūdas (3. tabulas pēdējā rinda) salīdzinot ar F-ABFC metodes šiem pašiem rezultātiem, ir ievērojami samazinājušās. Tas liecina par to, ka, atbilstoši teorijai, galējo labāko modeļu izvēles papildu validēšanas kopā

un šo modeļu vidējošanas dēļ, modelēšanas procesu patiešām ir iespējams padarīt stabilāku, kā arī robustāku pret intensīvu pārmeklēšanu, tādā veidā ievērojami uzlabojot prognozēšanu. Tomēr EF-ABFC metodei ir arī trūkums – kopumā modeļu būvēšanai ar izmantotajām datu kopām ar to bija nepieciešami aptuveni 7 reizes lielāki skaitļošanas resursi nekā modeļu būvēšanai ar F-ABFC. Neskatoties uz to, EF-ABFC metode kopumā tomēr ir ātrāka gan par SFS + AICC + AICC metodi (aptuveni 3 reizes), gan par SFS + AICC + CV metodi (aptuveni 12 reizes).

4. tabula

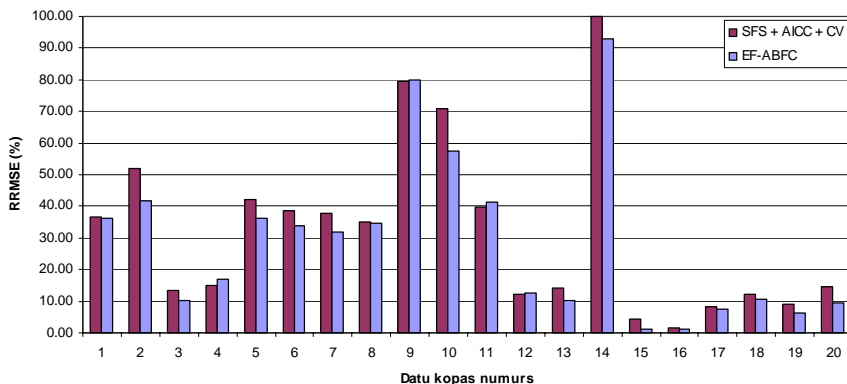
Salīdzināto regresijas modelēšanas metožu vidējie modelēšanai patērētie laiki

Metode	Vidējais laiks (pirmās 14 kopas)	Vidējais laiks (pēdējās 6 kopas)	Vidējais laiks (visas kopas)
Pilnais polin. + CV	-	~ 1 s	-
SFS + F-tests, $p = 1$	-	< 0.01 s	-
SFS + F-tests, $p = 2$	-	0.1 s	-
SFS + F-tests + CV	-	22 min	-
SFS + AICC, $p = 1$	0.7 s	< 0.1 s	0.5 s
SFS + AICC, $p = 2$	8.1 min	0.2 s	5.7 min
SFS + AICC, $p = 3$	-	3.8 s	-
SFS + AICC + AICC	1.3 h	57 min	1.2 h
SFS + AICC + CV	5.5 h	2.7 h	4.7 h
M5' RT	0.4 s	0.5 s	0.4 s
M5' MT	0.4 s	0.5 s	0.4 s
RBFNN	1.7 min	1.0 min	1.5 min
k-NN	< 0.01 s	< 0.01 s	< 0.01 s
MARS	3.3 min	54 s	2.6 min
LWP	-	25 min	-
CIPER + MDL	26 s	8.0 s	21 s
CIPER + AICC	15 min	1.9 min	11 min
F-ABFC	3.8 min	2.3 min	3.3 min
EF-ABFC	25 min	19 min	23 min

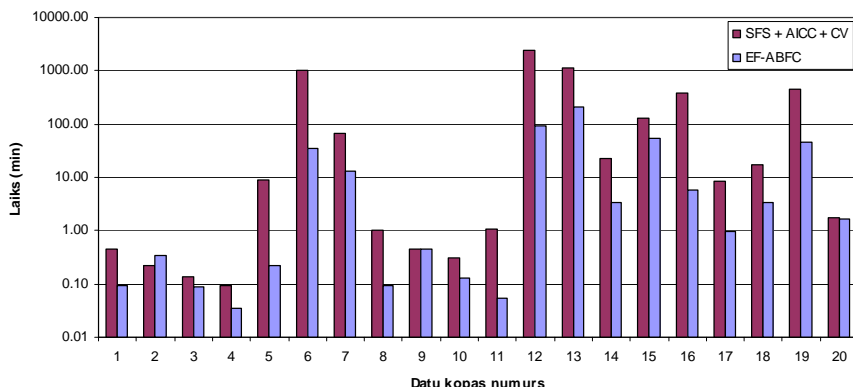
Kopumā, salīdzinot visas pilno polinomu, apakškopas atlasē pieejas un ABFC pieejas metodes, vidēji pa visām datu kopām vislabākā prognozēšanas spēja tika sasniegta, izmantojot EF-ABFC modeļu būvēšanas metodi. Pie tam tas tika izdarīts ievērojami īsākā laikā nekā ar apakškopas atlasē metodi SFS + AICC + CV, kura no apakškopas atlasē metodēm deva vislabākos rezultātus.

9. attēlā ir parādītas ar šīm divām metodēm iegūto modeļu prognozēšanas kļūdu diagramma, bet 10. attēlā ir parādīts modeļu būvēšanas procesā ar tām patērētais laiks (logaritmiskajā skalā). Kā redzams, EF-ABFC vairumā gadījumu pārspēj apakškopas atlasē metodi, ņemot vērā gan prognozēšanas kļūdu, gan patērēto laiku. 11. attēlā modeļu būvēšanai patērētais laiks ir parādīts lineārajā skalā, kur īpaši labi novērojams EF-ABFC ātrdarbības pārkums, būvējot modeļi 6., 12., 13. un 16. datu kopai. Piemēram, ar 12. datu kopu, modeļa būvēšana ar EF-ABFC prasīja aptuveni 1.6 stundas, bet ar SFS + AICC + CV – aptuveni 40 stundas, t.i., aptuveni 25 reizes vairāk. Visu šo četru datu kopu gadījumā augstais laika patēriņš ir saistīts galvenokārt ar lielo modelī iekļaut nepieciešamo bāzes funkciju skaitu augstas prognozēšanas spējas sasniegšanai, kas savukārt pie liela datu kopas

faktoru skaita nozīmē lielu izpildāmo pārmeklēšanas iterāciju skaitu. Laika patēriņu vēl vairāk palielina arī konkrētās kopas lielais faktoru skaits.



9. att. Apakškopas atlasē un EF-ABFC iegūto modeļu prognozēšanas kļūda katrā no salīdzināšanai izmantotajām datu kopām

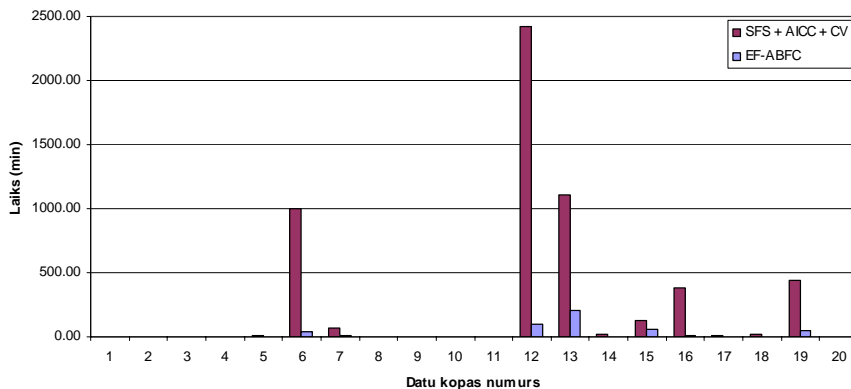


10. att. Apakškopas atlasē un EF-ABFC modeļu būvēšanas procesam nepieciešamais laiks (logaritmiska skala)

Veikto eksperimentu rezultāti apstiprina 4. nodaļā veiktās apakškopas atlasēšanas pieejas metožu un ABFC pieejas metožu efektivitātes teorētiskās analīzes rezultātus. Kopumā var secināt, ka, salīdzinājumā ar apakškopas atlasēšanas metodēm, apskatītajās datu kopās ABFC pieeja ļauj iegūt modeļus ar līdzīgu vai labāku prognozēšanas spēju īsākā laikā, pie tam bez nepieciešamības definēt pilno bāzes funkciju kopu vai modeļu maksimālo kārtu. Arī, salīdzinot ABFC metodes ar regresijas modelēšanas metodēm, kuras nesakņojas lineārā polinomu regresijā, tika secināts, ka EF-ABFC metode saglabā savu konkurētspēju.

Piektajā nodaļā tiek apskatīti arī divi konkrēti praktiski EF-ABFC metodes pielietojumi metamodelēšanas problēmu risināšanā Rīgas Tehniskās universitātes Materiālu un

Konstrukciju institūtā – lidmašīnu korpusu konstrukciju sabrukšanas metamodelēšana un stiklplasta klāja paneļu stiprības metamodelēšana. Tā ir atzīta par konkurētspējīgu ar citām metamodelēšanā populārām metodēm.



11. att. Apakškopas atlasēs un EF-ABFC modeļu būvēšanas procesam nepieciešamais laiks (lineāra skala)

Pirmajā pielikumā ir doti darbā lietotie vispārīgie regresijas modeļu precizitātes novērtēšanas mēri.

Otrais pielikums ir veltīts īsam matemātiskam mazāko kvadrātu metodes aprakstam.

Trešajā pielikumā ir dotas modeļu būvēšanas eksperimentu rezultātu tabulas.

GALVENIE DARBA SECINĀJUMI

Promocijas darbā tika iegūti šādi secinājumi:

1. Polinomu regresijas modeļu būvēšanas apakškopas atlasē pieejai ir trūkumi, kas saistīti ar netriviālo pilnās bāzes funkciju kopas vai maksimālās modeļu kārtas definēšanu un tās labākās apakškopas meklēšanai nepieciešamajiem eksponenciāli pieaugošajiem skaitļošanas resursiem.
2. Adaptīvā bāzes funkciju konstruēšanas pieeja var kalpot kā universāls karkass, uz kura pamata ir iespējams izstrādāt dažādas jaunas adaptīvas regresijas modeļu būvēšanas metodes ar šādām (no apakškopas atlasē metodēm atšķirīgām) īpašībām:
 - nav nepieciešams definēt pilno bāzes funkciju kopu vai izvēlēties modeļu maksimālo kārtu;
 - ir iespējams ģenerēt neierobežotas sarežģītības un kārtas modeļus;
 - modeļu būvēšanai nepieciešamo bāzes funkciju konstruēšana, un līdz ar to arī sarežģītības un kārtas noteikšana, tiek veikta pārmeklēšanas gaitā;
 - izveidojamo modeļu iespējamo funkcionālo formu paaugstinātās elastības dēļ tiek iegūta paaugstināta spēja piemēroties datos esošām sarežģītām nelineārām daudzfaktoru sakarībām;
 - salīdzinājumā ar apakškopas atlasē metodēm, tiek iegūtas samazinātas skaitļošanas resursu prasības.
3. Pie fiksēta faktoru skaita stāvokļu telpas zarošanās koeficients apakškopas atlasē pieejā palielinās eksponenciāli, bet ABFC pieejā – lineāri. Kopējie pārmeklēšanai nepieciešamie skaitļošanas resursi apakškopas atlasē pieejā palielinās eksponenciāli, bet ABFC pieejā – polinomiski.
4. Relatīvi pret apakškopas atlasē pieejas pārmeklēšanas ātrdarbību, ABFC pieejā pārmeklēšanas ātrdarbība pieaug līdz ar faktoru skaita un izvēlētas maksimālās kārtas palielināšanos un meklējamā modeļa sarežģītības samazināšanos. Pie tam, tā kā apakškopas atlasē pieejā modeļu maksimālā kārtā ir jāpiemeklē, jo tā iepriekš nav zināma, tad modelēšanas procesam nepieciešamie skaitļošanas resursi var vēl vairākkārtēji palielināties.
5. ABFC pieeja ir pielietojama ne tikai polinomu regresijā ar nenegatīvu pakāpju bāzes funkcijām, bet arī ar cita veida bāzes funkcijām, polinomu neironu tīklos, ar nelineārās regresijas modelēšanas metodēm, kā arī klasificēšanas problēmu risināšanā.
6. Veiktie regresijas modelēšanas metožu salīdzināšanas empīriskie eksperimenti ļāva secināt, ka praksē, salīdzinājumā ar apakškopas atlasē pieeju, ABFC pieeja ļauj iegūt līdzīgus vai labākus rezultātus, izmantojot ievērojami mazākus skaitļošanas resursus, vienlaicīgi neprasot piemeklēt modeļu maksimālo kārtu. Veiktajos eksperimentos modeļi ar vismazāko prognozēšanas kļūdu tika iegūti, ar ABFC pieejas speciālu gadījumu, izmantojot vidēji 12 reizes mazākus skaitļošanas resursus nekā labāko ar apakškopas atlasē pieeju iegūto rezultātu sasniegšanai.

Adaptīvās bāzes funkciju konstruēšanas pieejas izstrādes, novērtēšanas un pielietošanas procesā tika formulēti arī vairāki turpmāko pētījumu virzieni. Galvenie no tiem ir šādi:

- ABFC pārmeklēšanas telpas stāvokļu zarošanās koeficienta dinamiska kontrolēšana pārmeklēšanas gaitā ar augstas prognozēšanas spējas sasniegšanas mērķi minimālā laikā pie jebkura faktoru skaita (piemēram, izmantojot reaktīvās pārmeklēšanas principus);
- ABFC pieejas adaptēšana darbam ar negatīvu, daļveida pakāpju un cita veida bāzes funkcijām, nelineārās regresijas modeļiem, kā arī faktoru konstruēšanai klasificēšanas problēmās;
- ABFC speciāla gadījuma izstrāde, kura lietošana modeļu būvēšanā prasītu vēl mazākus skaitļošanas resursus, ņemot vērā faktoru skaitu, rezultātā iegūtā modeļa sarežģītību un apmācības kopas piemēru skaitu (piemēram, izmantojot ansambļa modeļu būvēšanai nepieciešamo aprēķinu apvienošanu, stāvokļu pārejas operatoru uzlabošanu, salikto operatoru izmantošanu, bāzes funkciju ortogonalizēšanu);
- ABFC pieejas pielietošana polinomu neironu tīklos;
- adaptīva modeļu būvēšana vairākām rezultatīvajām pazīmēm vienlaicīgi.

LITERĀTŪRA

- [Aha1991] Aha, D., Kibler, D. Instance-based learning algorithms. *Machine Learning*, Vol. 6, 1991, pp. 37-66.
- [Aha1994] Aha, D.W., Bankert, R.L. Feature selection for case-based classification of cloud types: An empirical comparison. *Proceedings of the AAAI'94 Workshop on Case-Based Reasoning*, 1994, pp. 106-112.
- [Aha1996] Aha, D.W., Bankert, R.L. A comparative evaluation of sequential feature selection algorithms. *Learning from Data*, Fisher, D., Lenz, H.J. (eds), New York, USA: Springer, 1996, pp. 199-206.
- [Aka1973] Akaike, H. Information Theory as an Extension of the Maximum Likelihood Principle. *Second International Symposium on Information Theory*, Petrov, B.N., Csaki, F. (eds), Budapest: Akademiai Kiado, 1973, pp. 267-281.
- [Aka1974] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 1974, pp. 716-723.
- [Bis1995] Bishop, C.M. *Neural networks for pattern recognition*. Oxford University Press, 1995
- [Blu1997] Blum, A.L., Langley, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, Vol. 97, 1997, pp. 245-271.
- [Chp2006] Chen, V.C.P., Tsui, K-L., Barton, R.R., Meckesheimer, M. A review on design, modeling and applications of computer experiments. *IIE Transactions*, Vol. 38, No. 4, 2006, pp. 273-291.
- [Clv1995] Cleveland, W., Loader, C. Smoothing by local regression: Principles and Methods (with discussion). *Computational Statistics*, 1995.
- [Cox1974] Cox, D.R., Snell, E.J. The choice of variables in observational studies, *Appl. Statist.*, Vol. 23, 1974, pp. 51-59.
- [Das1997] Dash, M., Liu, H. Feature selection for classification. *Intelligent Data Analysis. An International Journal*, Elsevier, Vol. 1, 1997, pp. 131-156.
- [Dre2006] Dreyfus, G., Guyon, I. Assessment methods. *Feature extraction: foundations and applications*, I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh (eds), Springer, 2006, pp. 65-88.
- [Egl1981] Eglajs, V. Approximation of data by multi-dimensional equation of regression. *Problems of Dynamics and Strength*, Vol. 39, Riga: Zinatne (in Russian), 1981, pp. 120-125.
- [Fan1996] Fan, J., Gijbels, I. *Local polynomial modelling and its applications*. London: Chapman and Hall, 1996
- [Fri1991] Friedman, J.H. Multivariate Adaptive Regression Splines (with discussion), *The Annals of Statistics*, Vol. 19, No. 1, 1991, pp. 1-141.
- [Fri1993] Friedman, J.H. *Fast MARS*. Tech. Report LCS110, Department of Statistics, Stanford University, 1993
- [Fuj2006] Fujita, K., Kounoe, Y. High-order polynomial response surface with optimal selection of interaction terms. *Proceedings of 11th AIAA/ISSMO multidisciplinary analysis and optimization conference*, Paper No. 2006-7054, Portsmouth, Virginia, USA, 2006
- [Gin1993] Ginsberg, M.L. *Essentials of Artificial Intelligence*. Morgan Kaufmann, 1993

- [Guy2003] Guyon, I., Elisseeff, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1157-1182.
- [Guy2006] Guyon, I., Elisseeff, A. An introduction to feature extraction, Feature extraction: foundations and applications, Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds), Springer, 2006, pp. 1-26.
- [Had2001] Hand, D.J., Mannila, H., Smyth, P. *Principles of data mining*, MIT Press, 2001, 578 p.
- [Has2001] Hastie, T., Tibshirani, R., Friedman, J. *The elements of statistical learning: Data mining, inference and prediction*, Springer, 2001, 552 p.
- [Jai1997] Jain, A., Zongker, D. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 2, 1997, pp. 153-158.
- [Jek2008] Jekabsons, G., Lavendels, J. Polynomial regression modelling using adaptive construction of basis functions. *Proceedings of IADIS International Conference, Applied Computing 2008*, Mondragon unibertsitatea, Algarve, Portugal, 2008, pp. 269-276.
- [Joh1994] John, G.H., Kohavi, R., Pflieger, K. Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning*, Cohen, W.W., Hirsh, H. (eds), SF, CA, USA: Morgan Kaufmann Publishers, 1994, pp. 121-129.
- [Kal2008a] Kalnins, K., Ozolins, O., Jekabsons, G. Metamodels in design of GFRP composite stiffened deck structure. *Proceedings of 7th ASMO-UK/ISSMO International Conference on Engineering Design Optimization*, Association for Structural and Multidisciplinary Optimization in the UK (ASMO-UK), Bath, UK, 2008, 11 p. (in print)
- [Kal2008b] Kalnins, K., Ozolins, O., Rucevskis, S., Eglitis, E., Wesolowski, M., Skukis, E., Ries, A., Rikards, R., Dzelzitis, K. Manufacture of the panels applying local damage and non-destructive inspection. *Improved Material Exploitation at Safe Design of Composite Airframe Structures by Accurate Simulation of Collapse*, COCOMAT, AST3-CT-2003-502723, Technical Report, RTU, Latvia, 2008.
- [Kit1978] Kittler, J. Feature set search algorithms, *Pattern Recognition and Signal Processing*, Sithoff and Noordhoff, Apthlen aan den Rijn, The Netherlands, 1978, pp. 41-60.
- [Koh1995] Kohavi, R., Sommerfield, D. Feature subset selection using the wrapper model: Overfitting and dynamic search space topology. *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, 1995, pp. 192-197.
- [Koh1997] Kohavi, R., John, G.H. Wrappers for feature subset selection. *Artificial Intelligence*, Vol. 97, 1997, pp. 273-324.
- [Lan1994] Langley, P. Selection of relevant features in machine learning. *Proceedings of the AAAI Fall Symposium on Relevance*, New Orleans, LA, USA: AAAI Press, 1994
- [Lar2006] Larose, D.T. *Data mining methods and models*, Wiley-IEEE Press, 2006, 344 p.

- [Lin2007] Lin, J.G. Modeling test responses by multivariate polynomials of higher degrees, *SIAM Journal on Scientific Computing*, Vol. 28, No. 3, 2007, pp. 832-867.
- [Mil2002] Miller, A. Subset selection in regression (2nd ed.), Chapman & Hall/CRC, 2002, 238 p.
- [Mol2002] Molina, L.C., Belanche, L., Nebot, A. Feature Selection Algorithms: A Survey and Experimental Evaluation. Proceedings of the International Conference on Data Mining (ICDM'02), IEEE Computer Society, Maebashi City, Japan, 2002, pp. 306-313.
- [Mye2002] Myers, R.H., Montgomery, D.C. Response surface methodology: process and product optimization using designed experiments (2nd ed). New York, USA: John Wiley & Sons, 2002
- [Pud1994a] Pudil, P., Ferri, F.J., Novovicova, J., Kittler, J. Floating search methods for feature selection with nonmonotonic criterion functions. Proceedings of the International Conference on Pattern Recognition, Vol. 2, IEEE, Los Alamitos, CA, 1994, pp. 279-283.
- [Pud1994b] Pudil, P., Novovicova, J., Kittler, J. Floating search methods in feature selection. *Pattern Recognition Letters*, Vol. 15, 1994, p. 1119-1125.
- [Qui1992] Quinlan, J.R. Learning with continuous classes. Proceedings of 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, 1992, pp. 343-348.
- [Rao1999] Rao, C.R., Toutenburg, H., Fieder, A. Linear models: least squares and alternatives, Springer, 1999, 427 p.
- [Ray1996] Rayward-Smith, V.J., Osman, I.H., Reeves, C.R., Smith, G.D. (eds) Modern heuristic search methods, Wiley, 1996, 314 p.
- [Reu2006] Reunanen, J. Search strategies. Feature Extraction: Foundations and Applications, Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (eds), Springer, 2006, pp. 119-137.
- [Rik1999] Rikards, R., Chate, A., Steinchen, W., Kessler, A., Bledzki, A.K. Method for identification of elastic properties of laminates based on experiment design, *Composites Part B*, Vol. 30, 1999, pp. 279-289.
- [Rik2003] Rikards, R. Identification of mechanical properties of laminates. Modern Trends in Composite Laminates Mechanics. CISM Courses and Lectures No. 448, Altenbach, H., Becker, W. (eds), Wien: Springer, 2003, pp. 181-225.
- [Rus2002] Russell, S.J., Norvig, P. Artificial intelligence: a modern approach (2nd edition), Englewood Cliffs, New Jersey: Prentice Hall, 2002
- [Sim2001] Simpson, T.W., Peplinski, J.D., Koch, P.N., Allen, J.K. Metamodels for computer-based engineering design: survey and recommendations, *Engineering with Computers*, Vol. 17, London: Springer-Verlag, 2001, pp. 129-150.
- [The2006] Theodoridis, S., Koutroumbas, K. *Pattern Recognition* (3rd ed), Academic Press, 2006, 856 p.
- [Tod2003] Todorovski, L., Dzeroski, S., Ljubic, P. Discovery of polynomial equations for regression. Proceedings of Sixth International Multiconference Information Society, Vol. A, Jozef Stefan Institute, Ljubljana, 2003, pp. 151-154.

- [Tod2004] Todorovski, L., Ljubic, P., Dzeroski, S. Inducing polynomial equations for regression. Proceedings of Fifteenth International Conference on Machine Learning, Berlin: Springer, 2004, pp. 441-452.
- [Wag2007] Wang, G.G., Shan, S. Review of metamodeling techniques in support of engineering design optimization. Journal of Mechanical Design, Vol. 129, Issue 4, 2007, pp. 370-380.
- [Wah1997] Wang, Y., Witten, I.H. Induction of model trees for predicting continuous classes. Proceedings of the Poster Papers of the Eighth European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, 1997, pp. 128-137.
- [Web2002] Webb, A.R. Statistical pattern recognition (2nd ed), John Wiley & Sons 2002, 496 p.
- [Wit2005] Witten, I.H., Frank, E. Data mining: practical machine learning tools and techniques with Java implementations (2nd ed), SF, USA: Morgan Kaufmann, 2005
- [Wol2006] Wolberg, J. Data analysis using the method of least squares, Springer, 2006, 250 p.
- [Zon1996] Zongker, D., Jain, A. Algorithms for feature selection: an evaluation. Pattern Recognition, Vol. 2, 1996, pp. 18-22.