Dima Alberg
Odsek za industrijski inženjering i upravljanje,
SCE - Shamoon College of Engineering,
Beer-Sheva, Israel

Inkrementalno izučavanje intervala primenom drveta regresije sa srednjom varijansom numeričkih tokova podataka

UDK: 005.521 ; 005.82
DOI: 10.7595/management.fon.2012.0013 (english version)

XIII Internacionalni Simpozijum SymOrg 2012, 05.-09. Jun 2012, Zlatibor, Srbija

U ovom radu predstavljamo novi model inkrementalnog učenja primenom drveta regresije koristeći numeričke tokove podataka struktuirane pomoću srednje varijanse. Predloženi MVIRT (Mean Variance Interval Regression Tree) algoritam pretvara kontinuirane vremenske podatke u dva statistička momenta u skladu sa vremenom koje je korisnik odredio i gradi model stabla regresije za procenu intervala predvidljivosti ciljne varijable. Glavna osobenost ovog algoritma jeste vremenski određen algoritam za indukciju inkrementalne varijanse koji se kombinuje sa novom rezolucijom vremena i mehanizmom za detekciju podataka koji odstupaju od uobičajenih. Rezultati tokova podataka u realnom vremenu pokazuju da se primenom MVIRT algoritma dobijaju precizniji modeli predviđanja koje je lakše tumačiti u poređenju sa drugim metodamama za serijsku obradu inkrementalnog modela drveta koji su danas u upotrebi.

Ključne reči: predviđanje, drvo regresione analize, inkrementalno učenje, analiza tokova podataka, predviđanje intervala

1. Uvod

Jedan niz momenata, na primer, podaci preneti putem senzora, geoprostorna lokacija, trenutne cene akcija i devizni kurs, električni signali i napon, vrednosti koje se mogu beležiti i posmatrati neprekidno u vremenu, naziva se *tok podataka*. Izvori tokova podataka su, između ostalog, meteorološki i finansijski podaci, kontrola mreže, web aplikacije, senzorske mreže, itd.

U slučaju regresije, atribut za koji se predviđaju vrednosti jeste numeričkog (stalno vrednovan i uređen), a ne kategorijskog tipa (vrednovan diskretnim vrednostima i neuređen). Ovaj atribut se može nazvati predviđenim atributom. Treba imati na umu da se regresija[7] može takođe posmatrati kao funkcija mapiranja, $Y=f(X)$, gde X predstavlja input, a autput je neprekidna ili uređena vrednost Y. Stoga ćemo problem regresije u tokovima podataka posmatrati kao problem učenja funkcije nekoliko ulaznih (input) atributa iz prethodnog toka podataka sa ciljem da što preciznije procenimo vrednost numeričkog ciljnog atributa u budućem toku podataka.

Klasični metodi regresionog drveta za predviđanje numeričke ciljne varijable u serijskoj obradi podataka nadograđuju se na kontrolisani pristup poznat kao „razdvoji i osvoji“. Kad je reč o predstavljanju poslednjeg čvorišta (lista), ovi metodi se mogu klasifikovati i sledeće kategorije:

- konstantna ili srednja vrednost
- model
- predviđanje intervala.

Glavna razlika između metoda jeste ta da za predstavljanje modela i predviđanja intervala, za razliku od pristupa primenom srednje vrednosti, koriste složenije modele predviđanja s ciljem da što preciznije predvide ciljnu varijablu. Efikasan algoritam regresionog modela drveta u serijskoj obradi obezbeđuje da konstrukcija i predviđanje modela mogu da se urade tako što će preciznost biti značajno unapređena pri unosu

novih informacija. Šta više, manje modele drveta, t.j., modele u kojima drvo ima manje račvanja lakše je i dobiti i tumačiti. U mnogim aplikacijama u stvarnim situacijama, u kojima postoji ogroman broj podataka, brza obrada i lakoća tumačenja regresionog modela drveta od jednake je važnosti kao i preciznost u predviđanju[6,10].

Prema Holmes i dr.[8], seriski-inkrementalni algoritmi nameću oštra ograničenja učenju primenom regresionog drveta. Prvo, model mora da bude inkrementalno uzrokovan. Drugo, instance koje sadrže vreme moraju da stižu istom brzinom. Treće, model mora da koristi konstantnu količinu memorije dok stvara jedan precizan i ažuriran model predviđanja, i to u svakom trenutku u vremenu. Stoga su, da bi prevazišli ova ograničenja, Bifet i dr.[3,4] uveli novu klasu grupe serijskih-inkrementalnih algoritama. Ovi algoritmi obično primenjuju grupu metoda padajućih prozora i razdvajaju tok podataka u tok nepovezanih serija, pri čemu svaka serija podataka može da se obrađuje onim redom kako je pristigla, primenom jednog od algoritama regresionog stabla o kojima ćemo podrobnije raspravljati u narednom poglavlju.

2. Analiza tokova podataka primenom metoda regresionog drveta

Iako su regresiona stabla već dobro proučena klasa metoda za učenje, malo je istraživanja urađeno u oblasti indukovanja inkrementalnog regresionog drveta. Problem postaje još ozbiljniji kada učenje ima posla sa neprekidnim tokovima podataka. To znači da onaj koji uči primenom regresionog drveta postaje u potpunosti inkrementalan i mora se ažurirati sa svakim novim unosom. U inkrementalnom sistemu učenja pojedinačni skupovi podataka pristižu prirodnim redosledom u vremenu. Svaki taj pojedinačni slučaj predstavlja jedan momenat u sistemu promena u okružanju u vremenu *t*. Pošto su tokovi podataka veoma dugi i često nemaju kraja, sistem se mora ažurirati instancu po instancu, skup po skup. Svaki skup se odbacuje čim se iskoristi za ažuriranje.

Potts[11,12] spaja serijsku i inkrementalnu verziju dva pravila razdvajanja modela linearnog drveta regresije pomoću dve unifikovane strukture Online-RD i RA. Predloženi modeli drveta grade se s vrha nadole, primenom jednog od dva statistička testa da bi se odredila tačka razdvajanja i odlučilo da li da se nastavi sa razdvajanjem. Ovi autori koriste Chow[4] test, standardni statistički test za homogenost između pod-uzoraka.

Alberg i Last[1,2] uvode MOPT (Mean Output Prediction Tree) algoritam za predviđanje intervala numeričkih ciljnih varijabli iz privremeno spojenih numeričkih podataka, gde je svaka instanca tako spojenih podataka predstavljena svojom srednjom vrednošću i varijansom. Predloženi algoritam se razlikuje od regresionih algoritama koji su danas u primeni po tome što raslojava ili razdvaja svaki input i autput na dva momenta, u skladu sa rezolucijom vremena i zato što može da identifikuje najbolju vremensku rezoluciju u predviđanju i time smanjuje grešku u predviđanju i stvara kompaktnije regresiono drvo bazirano na intervalima. Glavni nedostaci MOPT algoritma proizilaze iz činjenice da on ne koristi mehanizam za detekciju rezolucije eksplicitnog vremena te stoga nije pogodan kada se radi sa glomaznim tokovima podataka koji mogu da obuhvate i šeme promene u distribuciji i da zahtevaju prekomerno veliku memoriju i masivne mogućnosti obrade.

Ikonomovska i dr.[9] definišu FIMT-DD algoritam (Fast and Incremental Model Tree with Drift Detection) i predstavljaju ga kao naprednu adaptaciju FIMT (Fast and Incremental Model Tree) i FIRT (Fast and Incremental Regression Tree) algoritama koja primenjuje metod eksplicitne detekcije promene (DD) u slučajevima dinamičkih okruženja i distribucija koje se menjaju u vremenu. Glavna razlika između FIRT i FIMT algoritma jeste u činjenici da kod FIRT ne postoje linearni modeli u listovima. Prema ovim autorima, glavne prednosti FIMT-DD jesu da je on konkurentan „serijskim“ algoritmima u smislu preciznosti, omogućava detekciju lokalnih promena i omogućava da se izbegnu dodatni troškovi za ponovno građenje celog drveta kada su neophodne samo lokalne promene.

Uobičajeni nedostatak predstavljenih metoda jeste taj što oni ne mogu valjano da otkriju promene i da prilagode svoje drvo uz minimalni gubitak preciznosti. Jednostavan način da se izađe na kraj sa ovim problemom jeste da se izračuna svaka moguća tačka razdvajanja. Time zadatak postaje skup za izračunavanje, a to ima negativne posledice na skalabilnost algoritma. Ovo nije trivijalan problem i kao takav on zahteva primenu inkrementalnih algoritama za koje je karakteristično brzo rešavanje i brz vremenski odgovor, algoritama koji su u stanju da pravilno otkriju promene i prilagode svoje modele drveta uz minimalni gubital preciznosti.

3. Metodologija MVIRT

MVIRT algoritam predstavljen u ovom radu zahteva kontinuirane ukupne vremenske varijable predstavljene kao dva nepristrasna pokazatelja procene (prosek uzorka i varijansu) i proizvodi serijsko-inkrementalno drvo regresije za intervale za numeričku ciljnu varijablu *Y*.

U našem algoritmu, prosečna vrednost i varijansa svake varijable biće mapirane u univarijantnoj Mahalanobis distanci zasnovanoj na pomoćnoj kontrolnoj varijabli *M (·)* koja treba da reaguje na promene kod oba statistička momenta. Predloženi pristup omogućava nam da zapostavimo vrednosti koje odstupaju od normalnih i zbog kojih predviđanje može da izgubi na stabilnosti i da model bude opterećen overfiting efektom i da se tako značajno smanji veličina izgrađenog drveta.

$$X_i \sim \left\{\bar{x}_i(r);\ \hat{s}^2_{x_i}(r)\right\}, where\ i \epsilon (1, \dots, N) \tag{1}$$

Pretpostavimo da je svaka instanca *i* ulazne varijable X predstavljena sa dva ukupna estimatora srednje vrednosti i varijanse $\{\bar{x}_i(r)\}$ i $\left\{\hat{s}^2_{x_i}(r)\right\}$ za datu rezoluciju agregacije *r* merenja u vremenu. Uzmimo da su x_A $\{\bar{x}_i(r)\}$ i x_S $\left\{\overline{\hat{s}^2_{x_i}}(r)\right\}$ prosečne vrednosti uzorka za sve momente i da su V_A $\{\hat{s}^2(x_i(r))\}$ i V_S $\left\{\hat{s}^2\left(\hat{s}^2_{x_i}(r)\right)\right\}$ varijansa uzorka odgovarajućih nepristrasnih estimatora. Odgovarajuća kovarijansa uzorka između $\left\{\bar{x}_i(r);\ \hat{s}^2_{x_i}(r)\right\}$ označena je kao V_{AS} $\left\{\hat{s}^2\left(\bar{x}_i(r), \hat{s}^2_{x_i}(r)\right)\right\}$ a Mahalanobis distanca između dva merena statistička momenta ulazne varijable *X* izračunava se pomoću:

$$\begin{aligned} &M_i\left(\bar{x}_i(r);\ \hat{s}^2_{x_i}(r)\right) - \frac{N}{V_A V_S - V_{AS}} \cdot (*), where \\ &(*) = V_S \cdot (\bar{x}_i(r) - x_A)^2 + V_A \cdot \left(\hat{s}^2_{x_i}(r) - x_S\right)^2 - 2\sqrt{V_{AS}} \cdot (\bar{x}_i(r) - x_A) \cdot \left(\hat{s}^2_{x_i}(r) - x_S\right) \end{aligned} \tag{2}$$

Da bismo identifikovali vanredne vrednosti M_i potrebno je da odredimo distribuciju njegove verovatnoće. Predložene mere distance po nultoj hipotezi (koja obuhvata pretpostavku multivarijantne normalnosti varijable *X*) ima chi-square distribuciju sa dva stepena slobode i označava Mahalanobis multivarijantnu standardizovanu distancu između vrednosti dva prva momenta koja posmatramo. Na primer, ako prosečne vrednosti ulazne varijable zadrže vrednosti x_A i x_S, onda vrednosti *M(·)* treba da budu niže od i više od *0*, gde predstavlja gornju α procentnu tačku chi-square distribucije sa dva stepena slobode. Ako bar jedna od prosečnih vrednosti dobije neku novu vrednost, onda se verovatnoća da će statistički momenat preći granicu povećava. U algoritmu za indukovanje drveta, ograničenja intervala pouzdanosti metrike *M* distance izračunavaju se na sledeći način:

$$\begin{aligned} UCL(M_i) &= \frac{2(r-1)(n-1)}{rn-r-1} F_{\alpha/2, 2, rn-r+1} \\ LCL(M_i) &- \frac{2(r-1)(n-1)}{rn-r-1} F_{1-\alpha/2, 2, rn-r+1} \end{aligned} \tag{3}$$

gde *r* predstavlja vremensku rezoluciju agregacije.

4. Postupak račvanja drveta primenom MVIRT

U fazi razdvajanja u MVIRT algoritmu pretpostavljamo da imamo skup *n* trening momenata u datom čvorištu. Pseudo kod na slici 1 pronalazi najbolje mesto račvanja za predviđanje srednje vrednosti numeričke ciljne varijable. Ovaj postupak se primenjuje kod razdvajanja vrednosti bivarijantnih ulaznih varijabli gde je svaka varijabla predstavljena srednjom vrednošću uzorka *AVG(X)* i varijansom *VAR(X)* u skladu sa prethodno definisanom rezolucijom *r* vremenske agregacije.

Postupak grananja sastoji se iz tri glavna koraka. U prvom koraku izračunava se Mahalanobisova distanca za numeričku ulaznu varijablu *X* u svakom momentu (videti jednačinu 2) i sprovodi se postupak detekcije

vrednosti koje odstupaju od uobičajenih. Drugi korak sadrži logički mehanizam rezolucije inkrementalnog vremena koji povećava trenutnu rezoluciju vremena *TR* u slučaju kada svi momenti ulazne varijable odstupaju od normalnih vrednosti. Treba imati na umu da ukoliko je broj ovakvih vrednosti jednak broju trening momenata (instanci), onda algoritam zanemaruje datu ulaznu varijablu i prelazi na sledeću ili vraća drvo. U trećem, finalnom koraku odabira se najbolji estimator (uzorak srednje vrednosti ili varijanse) za ulaznu varijablu. U ovom koraku algoritam izračunava odnos apsolutnih razlika između vrednosti *MXY* i vrednosti estimatora *MAVG* i *MVAR* u poslednjem momentu razdvajanja X i odabira najbolji estimator čvorišta (Best_Contributor) koji će smanjiti odnos razlika na minimum.

MVIRT(α, TR, X, Y) postupak račvanja		
Ulazni Arg:	Definisan od strane korisnika , *α* Tekuća vremenska rezolucija, *r* Srednja varijansa ulazne varijable, *X* Srednja varijansa ciljne varijable, *Y*	
Izlaz:	Najbolja tačka račvanja za ulazni atribut *X*	
Glavnina pseudokoda:		
# Izračunati vektor Mahalanobis distance za ulaznu varijablu MX Za svaki momenat Do: MX = M (AVG(X),VAR(X)) (formula 2) Sledeće # Postupak detekcije i odbacivanja vrednosti koje odstupaju od normalnih {C} = nula Ako MX(*α*) ima vrednost koja odstupa od normalne (formula 2) Onda {C} + + # Prikupljanje momenata za podatke čija vrednost odstupa od normalnih Poslednji uslov		(1)

# Određivanje vremenske rezolucije 	
Ako je {C} prazan skup, Onda
# Preći na sledeću ulaznu varijablu 	
Vratiti se na MVIRT(α , TR, X, Y)
Ako vrednosti svih momenata varijable odstupaju od normalnih (<u>otkriveno je da koncept menja pravac</u>) Onda
# Povećati vremensku rezoluciju TR za sadašnju ulaznu varijablu 	
TR = *r* - -
Ako je TR prazan skup, Onda
Vratiti MVIRT drvo
Ili
Vratiti se na MVIRT(α, TR, X, Y)
Poslednji uslov
Poslednji uslov
		(2)

# Detekcija najbolje kontributivne varijable 	
Za svaki momenat u {C} Uraditi:
MY = M (AVG(Y), VAR(Y))
MXY = M (M(X), M(Y))
#Mahalanobis distancu između prosečnih vrednosti X i Y
MAVG = M (AVG(X), AVG (Y)) (formula 2)
#Mahalanobis distancu između varijansi X i Y
MVAR = M (VAR(X), VAR (Y)) (formula 2)
Sledeće
Racio najboljeg estimatora = Max(|MXY - MAVG |,| MXY - MVAR|)/ |MXY|
# Najbolja vrednost račvanja za sadašnju ulaznu varijablu X u vremenskoj rezoluciji TR 	
Vratiti Račvanje (TR; Najbolji estimator (Avg/Var) ; Najbolja vrednost račvanja)
		(3)

Slika 1. MVRT pseudo kod za kriterijum račvanja

5. Postupak konstruisanja lista drveta primenom MVIRT

U svakom završnom čvorištu MVIRT algoritam izračunava granice intervala predviđanja za odgovarajući list drveta sa nivoom pouzdanosti 1-α koju određuje korisnik primenom sledećih jednačina:

$$\begin{cases} \bar{y}_i \mp t_{1-\alpha/2,\, n_i-1} \cdot \hat{s}_{y_i}\sqrt{\left(1 + 1/n_i\right)}, & n_i \leq 30 \\ \bar{y}_i \mp z_{\alpha/2} \cdot \hat{s}_{y_i}\sqrt{\left(1 + 1/n_i\right)}, & n_i > 30 \end{cases} \tag{4}$$

gde $i \in (1,\ldots,n_l)$ predstavljaju momente listova na drvetu, a $\bar{y}_i$, $\hat{s}_{y_i}$ predstavljaju estimatore srednje i standardne devijacije na listu. Tako, kad stepen pouzdanosti bude jednak nuli (na primer, α = 100%), onda su odgovarajuće vrednosti distribucija $t_{0.5}$ i $z_{0.5}$ jednake nuli i MVIRT model transformiše prezentaciju intervala lista drveta u utvrđenu prosečnu vrednost ciljne varijable (kao po uzorku). Ovaj podatak je od velike koristi kad je reč o eksperimentalnom poređenju između MVIRT drveta i drugih algoritala drveta regresije kod kojih se procena vrši po tačkama.

6. Skup podataka za EL Nino

Tok podataka za El Nino može se dobiti na UCU KDD Archive (http://www.ics.uci.edu). Ovi podaci su sakupljeni pomoću Rešetke za tropsku atmosferu okeana (TAO rešetke) koja je projektovana u okviru programa izučavanja tropske globalne atmosfere okeana (TOGA program) (http://www.pmel.noaa.gov). TAO rešetka se sastoji od skoro 70 porinutih bova koje premrežavaju ekvatorsku oblast Pacifika i mere okeanografske i površinske meteorološke varijable koje su od kritičnog značaja za poboljšanu detekciju, razumevanje i predviđanje varijacija u klimi između godišnjih doba i između godina u tropskim predelima, posebno onih koje se odnose na cikluse El Nino/Južne oscilacije (ENSO cikluse). Ovaj tok podataka prikupljan je svakodnevno od marta 1980. do juna 1998. godine i sadrži 178.080 numeričkih momenata. Svaki momenat u ovom toku podataka sadrži sledeće numeričke atribute: datum, geografsku širinu, geografsku dužinu, zonske vetrove (zapadni <0, istočni>0), meridijanske vetrove (južni<0, severni>0), relativnu vlažnost, temperaturu vazduha, temperaturu vode na površini i temperature vodene mase do dubine od 500 metara. Geografska širina i dužina u podacima pokazale su da se bove kreću i stižu na različite lokacije. Podaci o vetrovima, kako zonskim tako i meridijanskim fluktuirali su između -10m/s i 10m/s. Vrednosti relativne vlažnosti u tropskoj oblasti Pacifika tipično su iznosile između 70% i 90%. Temperatura vazduha i temperatura površinskog sloja vode kretale su se između 20 i 30°Celzijusa. Ciljni atribut (za predviđanje) u toku podataka za El Nino jeste temperatura površine vode (SST) koja je po merenjima bila viša nego normalne temperature površine mora. U podacima neke vrednosti nedostaju. Kao što smo ranije napomenuli, sve bove nisu mogle da mere trenutne vrednosti atributa zato što su te vrednosti mogle da izostanu zbog nemogućnosti pojedinačne bove da ih izmeri. Operacja zamene nedostajućih vrednosti izvedena je postupkom interpolacije srednjih vrednosti susednih srednjih vrednosti obe vremenske serije. Konačno, da bismo mogli da procenimo predviđanje, čitav skup svih primera razložen je u skupove primera učenja i testiranja u proporciji 70:30.

Učinak MVIRT algoritma poredi se sa tri posojeća algoritma modela drveta odlučivanja koje koristi Rapid Miner, a koji su usaglašeni sa uključivanjem vremenskih serija M5P[13], Mp-Rules[14], Rep Tree14. Zbog ograničenja memorije i vremena bilo je veoma značajno proceniti sposobnosti svih algoritama da uče i inkrementalno i pravilno i da istovremeno konstruišu odgovarajući model drveta malih dimenzija. Stoga smo u svakom ogledu primenili mehanizam padajućeg prozora. Ovi mehanizmi u principu ne omogućavaju predviđanje intervala; stoga, da bi se izbeglo ovakvo ograničenje, koristili smo estimator srednje vrednosti padajućih prozora i tako napravili predviđanja po tačkama u našem eksperimentu komparativne procene. Konačno, u cilju poboljšanja skalabilnosti algoritma, uskladili smo M5P i Rep Tree sa mehanizmom procene celog pakovanja kakav je primenjen u Java API u WEKA paketu.

Rezultati prikazani na tabeli 1 pokazuju da pod uslovom prosečne greške u korenskom kvadratu srednje vrednosti (Average Root Mean Square Error (A[RMSE] i prosečne razjašnjene varijabilnosti (A[EV]), kriterijumi MVIRT i RETIS-M algoritmi postaju precizniji nego drugi predloženi algoritmi u smislu razlike u par-wise testa t-studenta. Označili smo znakom * one slučajeve gde je vrednost *p* razlike između MVIRT i drugih al-

goritama manja ili jednaka 5%. MVIRT algoritam je daleko efikasniji od drugih algoritama u smislu mere složenosti srednjeg troška (A[CCM]). Konačno, moramo da uzmemo u obzir i to da je naše predložene modele MVIRT drveta lakše tumačiti nego modele RETIS-M u smislu prosečne veličine drveta (A[TS]) (7 prema 23).

Tabela 1. Poređenje skupa potrebnih podataka za El Ninjo

Learner	A[RMSE]	A[TS]	A[CCM]	A[EV]
B-M5R	0,84*	7	1,01*	0,46*
B-M5P	0,83*	10	1,07*	0,47*
B-REPT	1,57*	5	1,69*	NA
M5 RLS	0,86*	7	1,03*	0,45*
M5P TR	0,84*	8	1,03*	0,46*
MVIRT	0,60	7	0,77	0,62
REPT	1,57*	3	1,64*	NA
RETIS	0,63	23	1,18*	0,60

Zaključak

U ovom radu smo predstavili MVIRT algoritam koji može da predvidi vrednosti numeričkih atributa glomaznih skupova podataka u vremenu. Predloženi algoritam razlikuje se od algoritama regresije koji se danas koriste po tome što svaku kontinualnu ulaznu osobenost razdvaja u skladu sa najboljom vrednošću koja doprinosi srednjoj varijansi, što u skupu trening podataka identifikuje vrednosti koje odstupaju od normalnih i konačno gradi kompaktnije drvo predviđanja intervala. Sprovedeni eksperiment ukazuje na to da predloženi MVIRT algoritam proizvodi preciznije i kompaktnije modele u poređenju sa algoritmima drveta regresije koji se trenutno koriste. Po našem mišljenju, predloženi algoritam predstavlja samo prvi korak ka čitavom skupu stvarno skalabilnih i brzih algoritama drveta regresije. Kad je reč o budućim pravcima istraživanja, ona se po našem mišljenju mogu odvijati u dva pravca. Prvo, naš opšti cilj jeste da stvorimo jedan on-line mehanizam koji bi precizno i snažno predviđao multi-r ciljne varijable, a koji se sa svoje strane može koristiti i za predviđanje velikih tokova podataka. Drugo, možemo da proučimo druge moguće analitičke metode za izbor tačke razdvajanja, čime bismo mogli da smanjimo složenost algoritma i što se tiče vremena i što se tiče prostora.

LITERATURA

[1] Alberg, D., Last, M., Neuman, R., & Sharon, A. (2009). Induction of Mean Output Prediction Trees from Continuous Temporal Meteorological Data. *2009 IEEE International Conference on Data Mining Workshops* (pp. 208 - 213). Miami, Florida, USA: IEEE Computer Society.

[2] Alberg, D., Last, M., & Kandel, A. (2011). Knowledge Discovery in Data Streams with Regression Tree Methods. *WIRES Data Mining and Knowledge Discovery* , 69-78.

[3] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavaldà, R. (2009). New Ensemble Methods for Evolving Data Streams. *In 15th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD'09).* Paris, France.

[4] Bifet, A., & Kirkby, R. (2009). *Data Stream Mining A Practical Approach.* COSI. Available at http://www.cs.waikato.ac.nz /~abifet /MOA.

[5] Chow, G. (1960). Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrica*, *28* (3), 591– 605.

[6] Friedman, J. (1991). Multivariate Adaptative Regression Splines. *In Annals of Statistics*, 1 - 19.

[7] Granger, C., & Newbold, P. (1986). *Forecasting in Business and Economics* (2nd Edition ed.). Academic Press.

[8] Holmes, G., Kirkby, R., & Bainbridge, D. (2004). *Batch Incremental Learning for Mining Data Streams.* Hamilton: University of Waikato, Department of Computer Science, New Zealand.

[9] Ikonomovska, E., Gama, J., & Dzeroski, S. (2010). Learning Model Trees from Evolving Data Streams. *Data Mining and Knowledge Discovery Journal*, Springer, (pp. 1 - 41).

[10] Krzanowski, W., & Hand, D. (2007). A Recursive Partitioning Tool for Interval Prediction. *In Proceedings of the ADAC (2007)*, *1*, pp. 241 - 254.
[11] Potts, D. (2004). Incremental Learning of Linear Model Trees. *Proceedings of the 21st International Conference on Machine Learning* (pp. 663 - 670). ACM.
[12] Potts, D., & Sammut, C. (2005). Incremental Learning of Linear Model Trees. *Machine Learning*, 5 - 48.
[13] Quinlan, J. (1993). Combining Instance-Based and Model-Based Learning. *In Proceedings of the 10th International Conference on Machine Learning* (pp. 236 - 243). Morgan Kaufmann.
[14] Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann.

Primljen: April 2012.
Prihvaćen: Jun 2012.

O autoru

Dima Alberg
Odsek za industrijski inženjering i upravljanje,
SCE - Shamoon College of Engineering, Beer-Sheva, Israel
E-mail: dimitria@sce.ac.il

Dr. Dima Alberg je predavač na Odseku za industrijski inženjering i menadžment na SCE Shamoon College of Engineering (Mašinski fakultet Šamon) u Izraelu. Osnovne fakultetske i magistarske studije u oblasti ekonomije i informatike završio je na Univerzitetu Ben Gurion, Negev. Na ovom univerzitetu takođe je odbranio doktorsku disertaciju iz oblasti projektovanja informacionih sistema. Oblasti njegovog naučnog interesovanja trenutno su poslovna inteligencija, analiza podataka i mašinsko učenje.