Data Mining: Evaluacija klasterovanja iz perspektive strima podataka

Olivera Janković
ORAO a.d.
Bijeljina, Republika Srpska, BiH
janolja@yahoo.com

Sažetak— **Problem klasterovanja u okviru data mininga iz perspektive strima podataka u mnogome predstavlja izazov u odnosu na tradicionalno data mining klasterovanje, dok validacija kvaliteta rezultujućeg klasteringa i odabir evaluacione metrike slovi za frustrirajući posao. U radu će biti prikazana evaluacija klasterovanja iz perspektive strima podataka i korištena evaluaciona metrika za validaciju eksperimentalnih performansi odabranih algoritama klasterovanja sa naglaskom na njihovu sposobnost da se nose sa šumom (*noise*) u sintetički generisanom strimu podataka, u okviru softverskog okruženja za masivnu online analizu (MOA).**

Ključne riječi- data mining; strim podataka; klasterovanje; evaluacija;

I. UVOD

Veliki kontinuirani tokovi podataka (*data stream*) su realnost i jedna od odlika sadašnjeg vremena sa velikom izvjesnošću povećane sveprisutnosti u kontekstu vremena koja tek dolaze. Izvori su različiti (socijalne mreže, razni sistemi logova podataka, senzorske mreže, finansijska tržišta,...), a sama narastajuća pojava strimova podataka najčešće se dovodi u usku vezu sa intenzivnim razvojem informaciono komunikacionih tehnologija, dok istraživanje ovako nastalih podataka (*data mining*) [1] može da predstavlja potencijalnu korist u širem spektru aplikativnih područja (industrija, nauka, poslovanje,...).

Imajući u vidu uobičajeno velike brzne kojim ovako nastali podaci pristižu čini ih zahtjevnim iz perspektive data mininga, zahtjevajući nove posebne tehnike i algoritme [2] koje mogu da se nose sa veoma striktnim ograničenjima prostora i vremena. Činjenica da podaci treba da budu istraženi u jednom prolazu (*one pass*) dovodi do toga da mnoge postojeće metode data mininga ne mogu biti primjenjene direktno na strim podataka. Tako na primjer široko korištene data mining operacije (klasifikacija, klasterovanje, ...) iz perspektive strimova podataka nose sa sobom velike izazove te to i dalje ostaje važna, nedovoljno istražena i otvorena tematika.

Klastering je jedan od važnih problema data mininga a data mining strimova podataka koristeći pristup klasteringa zauzima značajan obim novijih istraživanja [3], [4]. Pored odabira adekvatnih algoritama klasterovanja tu je i drugi ključni problem procjene performansi klasterovanja na razvijajućim (*evolving*) strimovima podataka. U okviru ovoga rada uz neophodan okvirni, teorijski kontekst biće prikazane i analizirane postignute pojedinačne i uporedne performanse CluStream i ClusTree algoritama klasterovanja sa naglaskom na njihovu sposobnost da se nose sa promjenama u podacima – konkretno sa šumom (*noise*) u sintetički generisanom strimu podataka (koji podržava simulaciju događaja klaster evolucije - spajanje i nestajanje klastera), te evaluaciona metrika korištena za validaciju performansi i poređenje klastering algoritama u okviru eksperimenata određenih postavki, korištenjem softverskog okruženja za masivnu online analizu (MOA) [5].

II. STRIM PODATAKA

Strim podataka se može formalizovano posmatrati kao poređana sekvenca tačaka podataka [6],

$$Y = \langle y1,y2,y3,... \rangle,$$

gdje indeks reflektuje redoslijed (bilo da je u pitanju eksplicitna vremenska oznaka, ili samo određena cijelobrojna vrijednost koja reflektuje redoslijed). Tačke podataka same po sebi su često jednostavni vektori u multidimenzionalnom prostoru, ali mogu takođe da sadrže nominalne/redne varijable, kompleksne informacije (npr. grafikoni) ili nestrukturisane informacije (npr. tekst).

Model strima podataka podrazumjeva da ulazni podaci nisu dostupni za nasumični pristup sa nekog memorijskog medija već stižu u formi kontinuiranog toka podataka. U biti model strima se razlikuje od standardnih relacionih modela u sledećem [7]:

- Elementi strima pristižu online;
- Redoslijed kojim elementi strima dolaze nije pod kontrolom sistema;
- Strimovi podataka su potencijalno neograničene veličine (beskonacni);
- Elementi strima podataka koji su obrađeni (procesuirani) su ili odbačeni ili arhivirani. Oni se ne mogu ponovno dohvatiti osim ako nisu smješteni u memoriju, koja je u pravilu mala u odnosu na veličinu strima;
- Obzirom na limitirane memorijske resurse i striktna vremenska ograničenja obrada podataka strima obično produkuje aproksimativne rezultate.

Sistemi strim podataka mogu konstantno da produkuju ogromne količine podatka (nrp. svaki senzor neke senzorske mreže ili logovi podataka mogu da šalju mjerenja svake sekunde). Ako se sagleda iz ugla skladištenja podataka, upravljanje i procesiranje, kontinuiran dolazak stavki podataka u višestrukim, brzim, vremenski varirajućim i potencijalno neogračenim strimovima, nameće nove izazove i probleme za istraživanje. Stoga je jasno da je obično teško izvodljivo prosto skladištenje podataka u okviru tradicionalnih sistema baza podataka s ciljem da se nad istim naknadno obave željene operacije. Umjesto toga, podaci strima se generalno moraju obraditi na online način kako bi se garantovalo da su rezultati ažurni (*up to date*), i da potencijalni upiti mogu dati odgovore sa malim vremenom kašnjenja.

III. DATA MINING: KLASTEROVANJE STRIMA PODATAKA

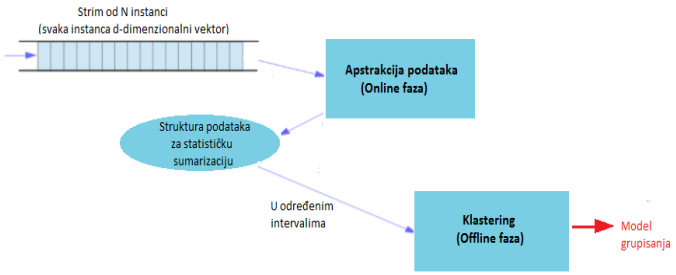
Za potrebe mašinskog učenja (*machine learning*) kao i istraživanja podataka (*data mining*) [8] koriste se podaci koji se čuvaju elektronski, na raznim vrstama i tipovima medija, pretraga je automatizovana korištenjem raspoloživih računarskih resursa Obzirom na napredak tehnika i metoda prikupljanja podataka, česti su slučajevi da se na podatke više ne gleda kao na statičke kolekcije, nego kao potencijalno vrlo veliki dinamički set, ili strim, dolaznih tačaka podataka. Pored klasifikacije [9] jedan od najčešćih zadataka data mining strima je i klastering [10], [11].

Klasterovanje je jedna od vrsta nenadgledanog učenja čiji cilj je pronalaženje obrazaca (šablona) u neobilježenim podacima (nema labele). Gruba definicija klasterovanja je da je klasterovanje proces organizovanja objekata u grupe čiji članovi su na neki način slični. Klaster je dakle kolekcija (grupa) objekata koji su "slični" međusobno, ali "različiti" od objekata koji pripadaju drugim klasterima [12**]**.

Strim se sastoji od n-torki (*tuple*) podataka koje je potrebno procesirati kako dolaze tako da je mining ovih strimova izazovan budući da distribucija podataka strima u osnovi može značajno da evoluira kroz vrijeme. Osim što postoji potreba da se nose sa problemom evoluirajuće distribucije, algoritmi strim klasteringa moraju da udovolje tehničkim zahtjevima, uključujući ograničeno vrijeme, ograničenu memoriju i da procesiranje strima obave u jednom prolazu [13]. Kako bi se reflektovale promjene u posmatranom strimu podataka klasterovanje u strim klasteringu konstantno se prilagođava.

Klastering strima ima dva tipa aplikacija: klasterovanje bazirano na atributima (*Attribute-based Clustering*) i klastering baziran na instancama (*Instance-based Clustering*). Cilj prvog tipa klasterovanja baziranog na atributima je pronaći grupe atributa koje se ponašaju na sličan način kroz vrijeme. Druga grupa aplikacije, korištena u ovom radu, sa druge strane klasteruje instancu kao cjelinu koja sadrži sopstvene atribute u d-dimenzionalnom prostoru. To zahtijeva dva odvojena koraka: faza apstrakcije podataka (poznata kao online faza) gdje se striming podaci rezimiraju na visokom nivou i faza klasteringa gdje je konačno grupisanje obezbjeđeno na osnovu komponenti isporučenih u online fazi.

Konceptualna reprezentacija ovog modela [14] može se vidjeti na Sl. 1, gdje je strim izvor za modul apstrakcije, online



Slika 1. Koncept modela klasterovanja baziranog na instanci

fazu i taj izlaz je nakon toga preusmjeren u dio za klastering, offline faza. Kao rezultat se dobije konačni model grupisanja.

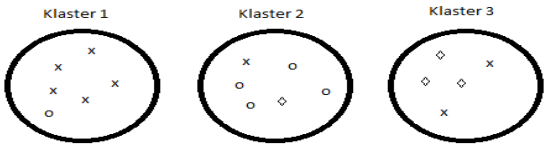
Pored već pomenutih ograničenja i nespornih izazova koje sa sobom nosi klasterovanje strima podataka tu je i određen broj, u tom kontekstu mogu da slove za manje probleme, kao što je problem određenih promjena u podacima strima koji se ogleda kroz pojavu šuma na primjer. U praksi mnogi strimovi podataka su kontaminirani određenim nivoom šuma koji se može pojaviti bilo pri procesu generisanja podataka strima ili prilikom njihovog prenosa, što u krajnjem otežava i utiče na tačnost potencijalnih prognoza.

Glavna značajka na kraju je svakako kvalitet rezultujućeg klasteringa, koji može biti mjeren evaluacionim mjerama, poznatim kao kriterijum (*criteria*), indeksi (*indices*), validacione mjere ili validacioni indeksi [15].

A. *Evaluaciona metrika*

Istraživanja evaluacionih mjera kada su u pitanju statički setovi podataka, na primjer tradicionalni klastering bez konteksta strima podataka, traju već nekoliko decenija. Po široko rasprostranjenom mišljenju klasterovanje samo po sebi se posmatra kao težak i subjektivan posao a validacija klastering rezultata često slovi i za najteži, frustrirajući dio klaster analize. Ovo važi za evaluaciju bez obzira da li je ili ne *ground-truth* (referentni podaci) dostupan u odnosu na koji se porede rezultati klasteringa. S druge strane korištenje ground-truth može striktno kategorizovati određene evaluacione mjere u interne (*internal*) [16] [17] i eksterne (*external*) [18] mjere evaluacije.

U osnovi interne mjere evaluacije uzimaju u obzir samo strukturu i osobine klastera. npr. njihova kompaktnost ili rastojanje međusobno. Eksterne mjere porede rezultujući klastering naspram referentnih podataka (ground-trouth). U literaturi [19] postoje kategorizacija postojećih evaluacionih mjera u kontekstu njihove pripadnosti internim odnosno eksternim mjerama evaluacije. U radu su korištene čistoća, preciznost, homogenost, kompletnost i SSQ mjere evaluacije.



Slika 2. Čistoća (Purity) kao eksterni evaluacioni kriterij za klastering podataka. Broj primjeraka većinske (*majority*) klase po klasterima je 5 (klaster 1), 4 (klaster 2) i 3 (klaster 3), a ukupan broj članova u svim klasterima je 17. Čistoća iznosi: (1/17)x(5+4+3)=0.705

Čistoća (*Purity*) [20] je jednostavna i transparentna evaluaciona mjera za određivanje kvaliteta klastera. Potrebno je za svaki klaster odrediti broj članova većinske, najbrojnije (*majority*) klase i njihov zbir za sve klastere podijeliti sa ukupnim brojem članova klasa u svim klasterima (Sl. 2). Inače visoku čistoću je lako postići kada je broj klastera velik (npr. čistoća će biti 1 ako svaki primjerak dobije svoj vlastiti klaster), stoga ne možemo koristiti čistoću u kontekstu vrednovanja kvaliteta grupisanja u odnosu na broj klastera.

Preciznost je važna mjera da se utvrdi efikasnost i tačnost bilo kog sistema za pronalaženje informacija. Ove mjera je omogućena i u okviru MOA softverskog okruženja, gdje F1_P predstavlja preciznost sistema.

Homogenost (*Homogeneity*) - Svaki klaster sadrži samo primjerke jedne klase. Kreće se od donje granice koja iznosi 0.0 do gornje granice koja je jednaka 1.0.

Potpunost ili kompletnost (*Completeness*) - Svi članovi date klase se dodeljuju na isti klaster. Donja granica je 0.0 i gornja granica je 1.0.

SSQ (*Sum of Square Distances*) u striming scenariju predstavlja zbir kvadrata rastojanja od tačaka podataka (objekata) do centra klastera. Ovdje je potrebno naglasiti da ova evaluaciona mjera nije normalizovana, zbog neograničene gornje granice.

IV. EKSPERIMENT EVALUACIJA KLASTEROVANJA STRIMA PODATAKA

A. *Radno okruženje*

Za testiranje je korišteno pomenuto MOA open source softversko okruženje za masovnu online analizu. Kompletan postupak, odnosno koncept toka postupka u okviru MOA okruženja se može uprošteno predstaviti sa odabirom i konfigurisanjem izvora podataka(*data feed*)/generatora, izborom i setovanjem karakteristika strim klastering algoritma i na kraju odabir same evaluacione metrike. Važno je pomenuti da nakon što je evaluacioni proces pokrenut postoji nekoliko opcija za analiziranje izlaza (*output*):

- Strim može biti zaustavljen u bilo kojem trenutku vremena i tekući (micro) klastering rezultati mogu biti prosleđeni kao set podataka (za npr. Weka explorer) za buduću analizu ili mining.
- Evaluaciona mjerenja koja su dobivena u konfigurabilnim vremenskim intervalima mogu biti smještena u .csv fajl kako bi se dobio željeni graf u offline režimu u okviru adekvatnog programa, proizvoljno odabranog.
- Klastering rezultati i pripadajuća mjerenja mogu biti vizuelizirani online u okviru okruženja za masivnu online analizu.

Mjere za analiziranje performansi klastering modela evaluiraju korektno raspoređene primjerke i unutrašnju strukturu rezultujućeg klasteringa (klasterovanja).

B. *Eksperimentalne postavke i rezultati*

Eksperimenti su izvršavani na Core i3-2328M CPU@ 2.20GHz računaru sa 4 GB memorije (Windows 8). Uokviru strim klastering okruženja korišteni su CluStream [21] algoritam koji održava statističke podatke o podacima pomoću mikro-klastera i samostalno-adaptivan (*self-adaptive*) ClusTree [22] klastering algoritam za istraživanje (mining) strimova podataka; korištenjem sintetički generisanog strima podataka posredstvom RBF (Radial Basis Function) generatora strima podataka.

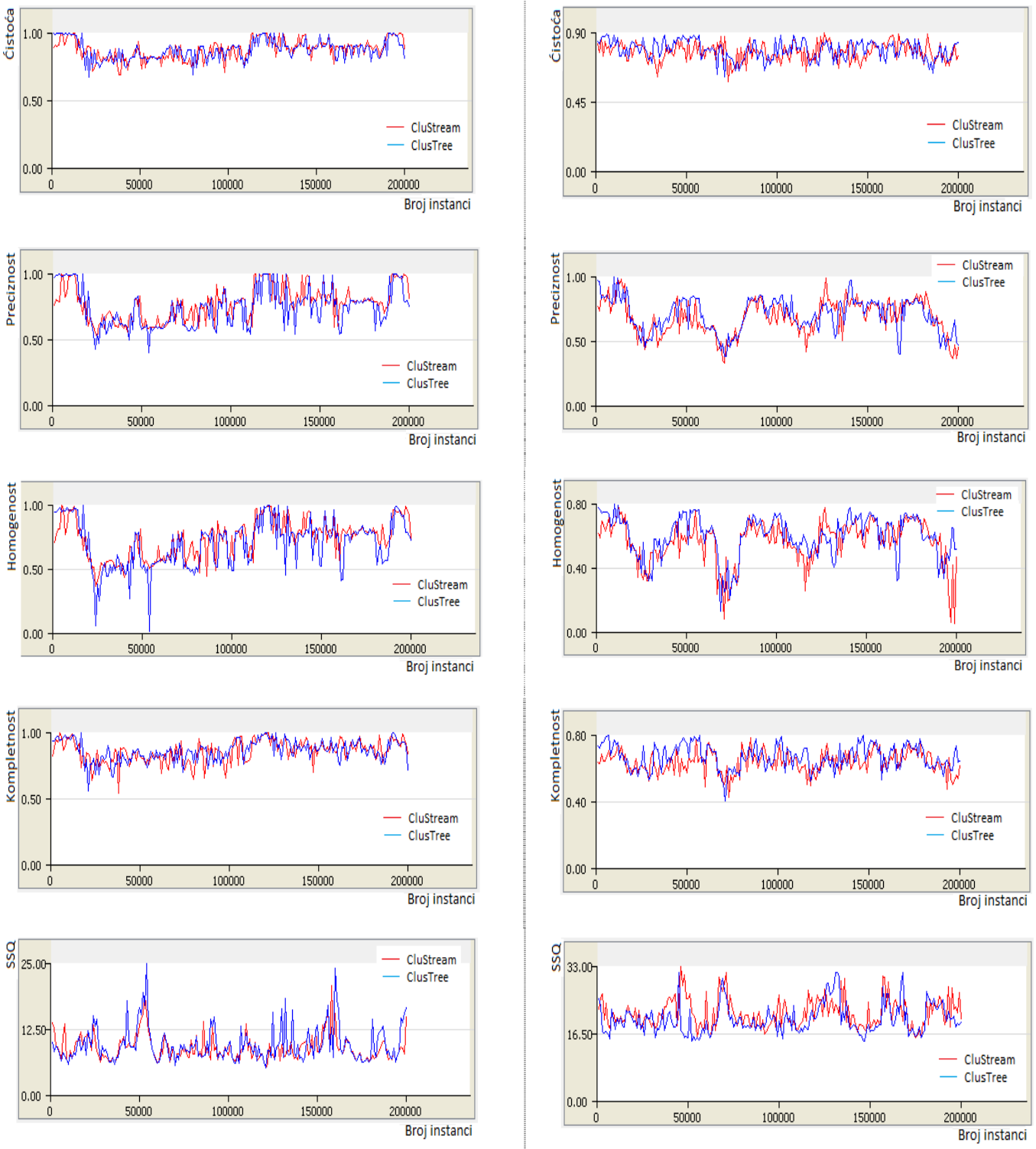
U okviru eksperimenata će se pratiti uticaj promjena, na sintetički generisanim podacima strima RBF generatora koji sadrži 200.000 tačaka kroz tri različite vrijednosti nivoa šuma, na postignute rezultate CluStream i ClusTree algoritama klasterovanja, sa ciljem poređenja performansi kako se svaki od njih pojedinačno i međusobno nose sa šumom, odnosno različitim nivoima šuma. Nivo buke će se kretati tako da će otprilike svaki 10-ti podatak (N=0.1), zatim svaki peti (N=0,2) i na kraju svaki treći podatak (N=0.33) biti slučajno (*randomly*) generisan.

TABELA I. EKSPERIMENTALNI REZULTATI CLUSTREAM I CLUSTREE ALGORITAMA KLASTEROVANJA, ZA RAZLIČTE NIVOE ŠUMA, KORIŠTENJEM PET MJERA EVALUACIONE METRIKE

		CluStream algoritam vrijednosti		ClusTree algoritam vrijednosti	
	Eval. mjere	Finalna	Srednja	Finalna	Srednja
N=0.1	F1-P (Preciznost)	0.80	0.77	0.74	0.74
	Čistoća (Purity)	0.87	0.87	0.81	0.86
	Homogenost	0.71	0.71	0.70	0.67
	Kompletnost	0.81	0.83	0.70	0.84
	SSQ	14.30	8.60	16.06	9.23
N=0.2	F1-P(Preciznost)	0.55	0.61	0.50	0.61
	Čistoća (Purity)	0.77	0.78	0.82	0.79
	Homogenost	0.52	0.52	0.25	0.52
	Kompletnost	0.78	0.72	0.68	0.74
	SSQ	13.73	15.59	20.14	15.08
N=0.33	F1-P(Preciznost)	0.40	0.62	0.42	0.63
	Čistoća (Purity)	0.70	0.72	0.78	0.74
	Homogenost	0.46	0.55	0.51	0.59
	Kompletnost	0.59	0.61	0.62	0.64
	SSQ	19.66	21.09	19.00	19.63

Rezultati postignuti u okviru ovih eksperimentalnih postavki dati su Tabelom I, dok se čistoća, preciznost, i ostale evaluacione mjere, predstavljene pojedinačno, u formi uporednih grafova za korištene CluStream i ClusTree algoritme, u odnosu na broj instanci, za vrijednosti šuma N=0.1 i N=0.33, u okviru lijevog dijela slike Sl.3.a), odnosno desnog dijela Sl.3.b) respektivno. Izabrane su vrijednosti za ove dvije od tri vrijednosti šuma jer su reprezentativne u kontekstu gradacije postignutih rezultata klastering modela.

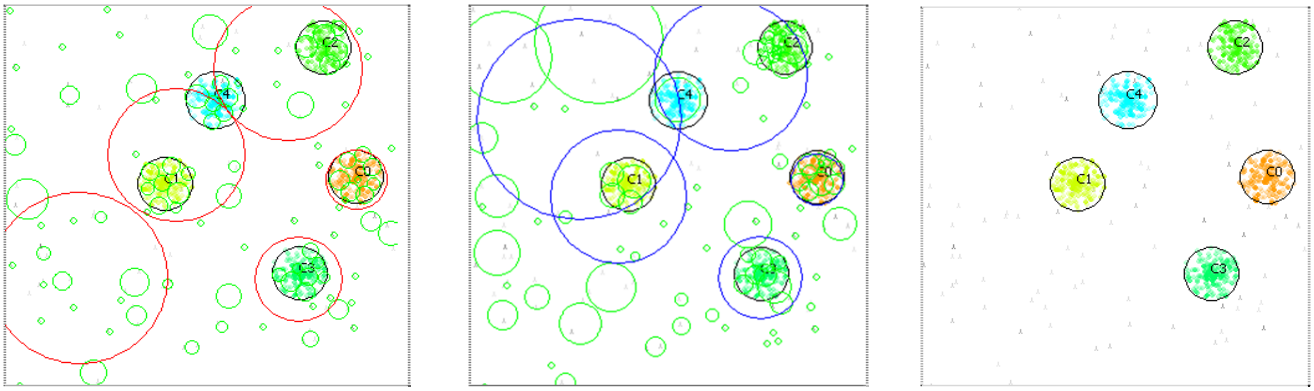
Naime, ako se posmatra čistoća, kao prva po važnosti izabrana evaluaciona mjera može se zaključiti na osnovu rezultata iz Tabele I da su finalne vrijednosti čistoće, nakon 200.000 tačaka, za CluStream algoritam, opadale sa porastom šuma: od 0.87 (za N=0.1), preko 0.77 (N=0,2) do 0.70 (N=0.33). U slučaju ClusTree algoritma, situacija je nešto drugačija i može se na osnovu rezultata datih tabelom reči da tu nije došlo do značajnijih odstupanja sa povećanjem šuma



a) N = 0.1

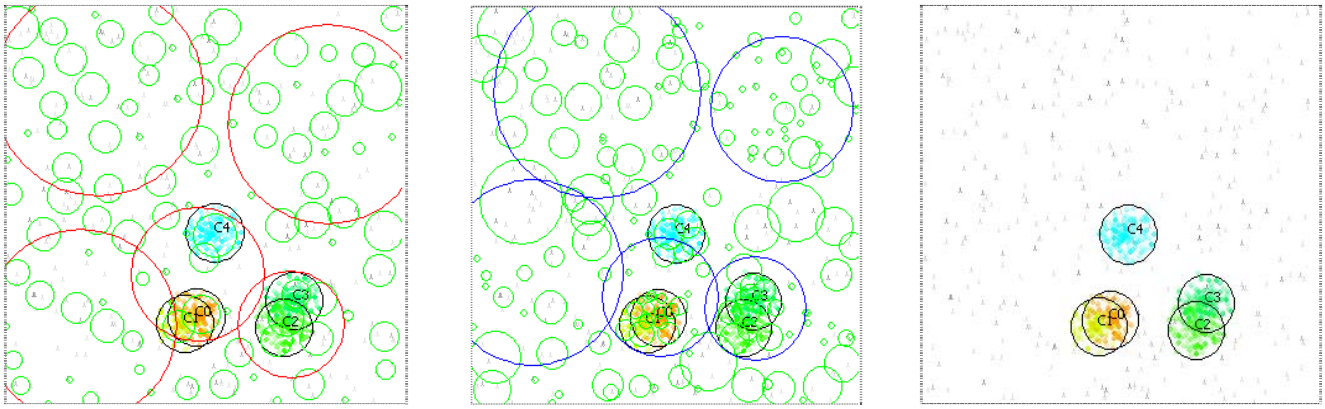
b) N = 0.33

Slika 3. Prikaz vrijednosti korištenih evaluacionih mjera (Čistoća, Preciznost, Homogenost, Kompletnost, SSQ) u formi uporednih grafova, za korištene CluStream i ClusTree algoritme klasterovanja, u odnosu na broj instanci sintetički generisanih podataka strima (200.000 instanci) pri: a) nivo šuma N=0.1 i b) nivo šuma N=0.33



a) CluStream algoritam b) ClusTree algotitam c) Ground-trouth

Slika 4. Vizuelan prikaz rezultata klasterovanja za 200.000 instanci i nivo šuma N=0.1



a) CluStream algoritam b) ClusTree algotitam c) Ground-trouth

Slika 5. Vizuelan prikaz rezultata klasterovanja za 200.000 instanci i nivo šuma N=0.33

(0.81, 0.82, 0.78) u poređenju sa CluStream algoritmom. S druge strane u međusobnom poređenju CluStream je postigao najbolji rezultat (0.87) u cjelosti posmatrano i bolje rezultate za svaku od vrijednosti šuma u odnosu na ClusTree algoritam.

Što se tiče procjene upotrebom evaluacione mjere preciznost može se zaključiti da su se oba algoritma posmatrana pojedinačno ponašala slično, imali su značajniji pad finalnih vrijednosti preciznosti sa porastom vrijednosti šuma, dok je u međusobnom poređenju za sve nivoe šuma bio (ponovno) bolji CluStream algoritam.

Homogenost je bila prilično ujednačena, uporedno posmatrana za N=0.1, pri čemu je istovremeno i najveća vrijednost homogenosti (najbolja) za CluStream za najmanji šum, ukupno posmatrano. Vidljivo je takođe da vrijednosti homogenosti padaju sa povećanjem šuma, što je još jedan od pokazatelja nepovoljnog uticaja šuma kada su u pitanju korišteni algoritmi. U tom kontekstu slična situacija je i kada je u pitanju evaluaciona mjera kompletnost, čije vrijednosti opadaju sa povećanjem nivoa šuma.

Prilikom procjene algoritama korištenjem SSQ evaluacione mjere potrebno je znati da su manje vrijednosti bolje te u tom kontekstu analizirati postignute rezultate.

Na Sl.4 i Sl.5 se može vidjeti i rezultujući vizuelni prikaz klasterovanja za 200.000 instanci, za prethodno izabrane vrijednosti šuma u kontekstu gradacije postignutih rezultata klastering modela, N=0.1 i N=0.33 respektivno, za ClusTree i CluStream algoritme kao i referentni, ground-trouth.

ZAKLJUČAK

U radu su evaluirani eksperimentalni rezultati korištenjem pet mjera evaluacione metrike. Ekperimentalni rezultati i analiza istih generalno pokazuju, sa manje/više različitim odnosima i odstupanjima, trend pada performansi korištenih ClusTree i CluStream algoritama sa povećanjem nivoa šuma, bez obzira na korištenu mjeru evaluacije, što potencijalno ostavlja prostor za nova rješenja kada je u pitanju pojava šuma u podacima strima na primjer.

Evaluacija mininga strima podataka, obzirom na njihovu sveprisutnost i tendenciju rasta iste, predstavlja važan zadatak, za čija rješenja možemo reći da su još uvjek u ranoj fazi razvoja, te je stoga realno za očekivati nova poboljšana rješenja kako u domenu mogućnosti algoritama klasteringa tako i domenu adekvatnosti evaluacione metrike.

LITERATURA

[1] A.Bifet, G. Holmes, R. Kirkby. and B. Pfahringer, DATA STREAM MINING A Practical Approach, The University of Waikato, COSI, 2011

[2] C.C. Aggarwal, Data Streams Models and Algorithms, Springer Science+Business Media, LLC, 2007

[3] A. Amini, T. Ying Wah, and H. Saboohi, "On Density-Based Data Streams Clustering Algorithms: A Survey ",Journal of computer science and technology, 2014, 29(1), pp. 116–141.

[4] Z.R. Hesabi, T. Sellis, and X.Zhang, "Anytime Concurrent Clustering of Multiple Streams with an Indexing Tree" , JMLR: Workshop and Conference Proceedings 41:19–32, 2015

[5] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cs.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010

[6] M. Hahsler, M. Bolanos, and J. Forrest. "Introduction to stream: An Extensible Framework for Data Stream Clustering Research with R", Journal of Statistical Software, 2015.

[7] J. Beringer and E. Hullermeier, "Online Clustering of Data Streams" Technical report, University of Marburg, 2003.

[8] O. Janković, "Mašinsko učenje: Evaluacija performansi klasifikatora u kontekstu ograničene raspoloživosti podacima", XLI Simpozijum o operacionim istraživanjima – SYM-OP-IS 2014., Zbornik radova, str. 232-237.

[9] O. Janković, "Data Mining: Evaluacija klasifikacije iz perspektive strima podataka", XXII naučna i biznis konferencija YU INFO 2016, prihvaćen za objavljivanje

[10] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "On clustering massive data streams: A summarization paradigm", In Charu C.Aggarwal, editor, Data Streams: Models and Algorithms, chapter 2. Springer, New York, 2007

[11] J. A. Silva, E.R. Faria, R.C. Barros, E.R. Hruscheka, A. Carvalho, and J.P. Gama, "Data Stream Clustering: A Survey", ACM Computing Surveys (CSUR), 2013, vol 46 , pp 13:1–13:31.

[12] P-N. Tan, M. Steinbach, and V.Kumar, "Introduction to Data Mining", Addison-Wesley, 1st ed., 2005

[13] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2002, New York, pp. 1–16.

[14] J.A. Merino, "Streaming Data Clustering in MOA using the Leader Algorithm", Master thesis, Universitat Politecnica de Catalunya, 2015

[15] H. Kremer, P.Kranen, T. Jansen, T. Seidl A. Bifet, G.Holmes, B. Pfahringer "An Effective Evaluation Measure for Clustering on Evolving Data Streams", 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp.868-876.

[16] G. W. Milligan, "A monte carlo study of thirty internal criterion measures for cluster analysis," Psychometrika, vol. 46, no. 2, pp. 187–199, 1981

[17] M. Hassani, T. Seidl, "Internal Clustering Evaluation of Data Streams", Trends and Applications in Knowledge Discovery and Data Mining Volume 9441 of the series Lecture Notes in Computer Science, November 2015, pp 198-209

[18] J. Chen, "Adapting the right measures for k-means clustering," in ACM KDD, 2009, pp. 877–884.

[19] P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes and B. Pfahringer,. "Clustering Performance on Evolving Data Streams: Assessing Algorithms and Evaluation Measures within MOA", ICDM Workshops, 2010, pp. 1400–1403.

[20] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[21] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29, VLDB '03, pp. 81–92, 2003.

[22] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The clustree: indexing micro-clusters for anytime stream mining," Knowledge and Information Systems, vol. 29, no. 2, pp. 249–272, 2011.

ABSTRACT

The problem of clustering in data mining from the perspective of stream data in many ways represents a challenge to the traditional data mining clustering. Validation of the quality of the resulting clustering and selection of evaluation metrics is considered to be a frustrating business. This paper will be presented from the perspective of the evaluation of clustering data stream. and used evaluation metrics for validation of experimental performance of selected clustering algorithms. The emphasis is on their ability to cope with the noise in a synthetic stream of data that is generated within the software environment for massive online analysis (MOA).

DATA MINING: EVALUATION OF CLUSTERING FROM THE PERSPECTIVE OF DATA STREAM

Olivera Janković