Примена анализе кластера у понашању учесника у саобраћају у вези са коришћењем система безбедности и мобилних телефона

Стефан Шошић
Факултет техничких наука, Чачак
Информационе технологије, Инжењер, 2020/2021.
stefan.sosic@ftn.kg.ac.rs
Ментор рада: др Марија Благојевић

Апстракт— Овај рад представља анализу кластера која се односи на понашање учесника у вези са коришћењем безбедносних система и мобилних телефона. Подаци о понашању у саобраћају преузети су са портала отворених података у Србији. Примењене су три врсте анализе кластера: хијерархијско кластеровање, кластеровање Бајесовим критеријумом информација (БИЦ) и моделно кластеровање. Добијени резултати указују на различите могућности коришћења ове три методе кластеровања у области саобраћаја.

Кључне речи – кластеровање, анализа, саобраћај, прекршаји

1 Увод

Упоредно са развојем друштва, развија се и саобраћај, као и комуникација у саобраћају. Основни фактор развоја сваког друштва је саобраћај. Ниво развоја саобраћаја користи се за мерење нивоа развоја одређеног друштва. Саобраћај у Србији је од изузетног значаја због локације земље на раскршћу Балкана. У области законодавства, у Србији су усвојене следеће регулативе:

1. Закон о безбедности у саобраћају – Службени гласник Републике Србије број: 41/2009,53/2010, 101/2011, 32/2013. [1]
2. Стратегија транспорта Републике Србије, 2015 – 2025. [2]

У ери отворених података, подаци о саобраћају у Србији добили су и свој простор. Посебан одељак српског портала отворених података [3] посвећен је јавној безбедности и он је био извор података који се користе за анализу кластера у овом раду.

Праћење и анализа саобраћаја играју кључу улогу у подизању нивоа транспорта робе и путника. Статистика и показатељи који карактеришу саобраћај су бројни и често је њихова колекција и формирање база података ограничена њиховом доступношћу. Примена савремених статистичких и математичких метода у вредновању саобраћаја омогућава свеобухватну анализу која указује на велики број индикатора, као и велику количину података.

Циљ рада је груписање учесника у саобраћају према средини у којој је почињена највећа количина саобраћајних прекршаја.

2 Методологија кластеровања

Техника анализе података која је примењена за решавање проблема истраживања је кластеровање. Да би разумели технику кластеровања, прво треба објаснити термин кластер. Кластер се односи на групу објеката који припадају истој класи. То значи да су слични објекти груписани у један кластер, а различити објекти у други кластер. На основу тога, кластер објеката са подацима може бити представљен као једна група. Процес прављења групе објеката података по сличности назива се кластеровање. Према [4], кластеровање је техника класификације без надзора у анализи шаблона. Главна предност ове технике је што је прилагодљива променама и помаже да се одвоје корисне карактеристике које разликују разноврсност група. У овом раду, учесници у саобраћају су груписани према најчешћој локацији где су починили прекршаје, у насељу, ван насеља или на аутопуту.

Важна ствар коју треба узети у обзир приликом избора алгоритма кластеровања је да ли се алгоритам скалира у односу на скуп података који се користи за кластеровање. Алгоритам би требало да има добре перформансе и ефикасност с обзиром да скуп података који се користи за груписање може да садржи огромну количину података. [5]

Један од једноставнијих алгоритама учења који решава проблем кластеровања је “K-means” и он се може применити на ове резултате. Идеја је да се дефинишу к центри за сваки кластер. Различита локација к центара

даје другачији резултат. Након тога се креира петља и као резултат тога, к центри мењају локацију корак по корак док не престану да се померају. Овај алгоритам има циљ да умањи грешку у квадрату објективне функције:

$$J(V) = \sum_{i=1}^{C} \sum_{j=1}^{Ci} (||Xi - Vj||)^2 \quad (2.1)$$

где параметри функције представљају следеће:

||Xi – Vj|| - Еуклидска (енг. Euclidean) дистанца између Xi и Vj

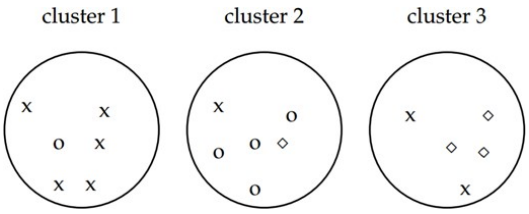
Ci – број тачака података у кластеру на i-тој позицији

C – број центара у кластеру

Када се израчуна први центар кластера, следећи мора бити поново израчунат помоћу следеће функције:

$$Vi = (\frac{1}{Ci}) \sum_{j=1}^{Ci} Xi \quad (2.2)$$

Помоћу ове функције поново се израчунава растојање између сваке тачке података и новог добијеног центра кластера. Ако ниједно место података није поново додељено, процес се зауставља. [6]



Слика 1. Кластери након завршетка процеса израчунавања центра кластера

Технике кластеровања које су примењене на скуп података индикатора понашања учесника у саобраћају у Србији су:

- хијерархијско кластеровање;
- кластеровање засновано на Бајесовом информационом критеријуму;
- кластеровање засновано на моделима;

2.1 Хијерархијско кластеровање

Хијерархијско кластеровање се може поделити на две врсте хијерархијских стратегија анализе кластера: агломеративне и дељиве.

Хијерархијско *агломеративно кластеровање* (енг. HAC - Hierarchical agglomerative clustering), познато и као приступ одоздо ка горе, информативније је од неструктурираног скупа кластера враћених равним кластеровањем. Равно кластеровање представља начин груписања када научник говори програму у колико категорија треба груписати податке. Док код хијерархијског кластеровања је другачије. Програму је дозвољено да одлучи колико кластера ће створити на основу сопствених алгоритама. Алгоритми који се користе "HAC" третирају сваки податак као посебан кластер на самом почетку, а затим узастопно упарују парове кластера док се сви кластери не обједине у један кластер који садржи све податке.

Са друге стране, *кластеровање дељењем* познато је као приступ од врха на доле. Захтева метод за поделу кластера који садржи целе податке и наставља се тако што се кластери поново деле док појединачни подаци не буду подељени на појединачне кластере. [7]

На основу наведених података, код HAC за почетак се мора израчунати Еуклидска (енг. Euclidean) дистанца:

$$D(Xi, Xj) \quad (2.3)$$

Xi – Xj представља основно растојање између било која два елемента X, и минимално растојање за дефинисање под-сета растојања:

$$\Delta(Xi, Xj) = \min(X) \in X, Y \in Xj \; D(x, y) \quad (2.4)$$

2.2 Кластеровање засновано на Бајесовом информационом критеријуму

Кластеровање засновано на Бајесовом информационом критеријуму (енг. BIC), које је предложио Schwarz [8], према [9] претпоставља да ће изабрати један међу скупом кандидирајућих модела M = M1, M2... Mm за представљање датог скупа података D = D1, D2....Dn.

„BIC" модел од Mi представљамо следећом формулом:

$$BIC(Mi) = logP\,(D1, D2 \ldots DN \mid Mi) - 1/2di\, log\, N \quad (2.5)$$

где, Di је број независних параметара у моделу Mi и P(D1,D2,...,Dt| Mi) је максимална вероватноћа модела.

2.3 Кластеровање засновано на моделима

Различити алгоритми за кластеровање имају различите објективне функције, али општа идеја је да се смањи растојање између објеката у истом кластеру уз максимално растојање између објеката у различитим кластерима. Минимизација унутра кластерске дистанце се такође може посматрати као минимизација растојања између

сваког податка Xi и кластера значи Cj. С обзиром на скуп кластера, Cj је очекивана квадратна сума процењених грешака која се може израчунати на следећи начин:

$$E\left(\sum_{j=1}^{k}\sum_{j\in c}\left\|Cj - Xi\right\|^{2}\right) = \sum_{j=1}^{k}\sum_{i\in Cj}\int \|Cj - Xj\|^{2} f(Xi)dXi \quad (2.6)$$

где || . || је метрика растојања између тачке података Xi и средина кластера Cj.

Средина кластера се добија на следећи начин:

$$Cj = E\left(\frac{1}{|Ci|}\sum_{i\in Ci} X\right) = \frac{1}{|Ci|}\sum_{i\in Cj}\int Xi\ f(Xi)\,dXi \quad (2.7)$$

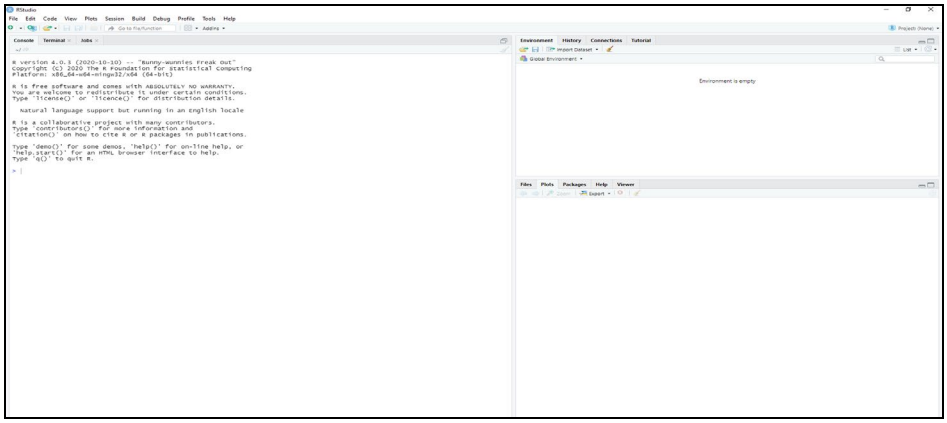
3 Р ПРОГРАМСКИ ЈЕЗИК И РАЗВОЈНО ОКРУЖЕЊЕ RSTUDIO

Р је статистички програмски језик који је брзо стекао популарност у многим научним областима. Развили су га George Ross Ihaka и Robert Clifford Gentleman као имплементација отвореног кода програмског језика „С“. Р је такође назив софтвера који користи овај језик за статистичко рачунање. Уз огромну заједницу подршке на мрежи и наменске пакете који пружају додатну функционалност за практично било коју апликацију и поље студија, тешко да постоји нешто што не можете да урадите у Р-у.

RStudio је доста сличан осталим статистичким програмима попут Минитаба или СПСС-а, док је главна разлика у томе што Р нема графички кориснички интерфејс, што значи да нема дугмади за клик и нема падајућих менија. Р се може у потпуности покренути куцањем команди у текстуални интерфејс. Ово може изгледати мало застрашујуће, али то значи и пуно већу флексибилност, јер се за своје анализе не ослањате на унапред утврђени скуп алата.

Р програмски језик сам по себи нема графички интерфејс, али већина људи комуницира са Р-ом путем графичких платформи које пружају додатну функционалност. Користићемо програм под називом RStudio као графички предњи крај Р-а, тако да можемо приступити нашим скриптама и подацима, пронаћи помоћ и прегледати графиконе и излазе на једном месту.

RStudio је један од најпопуларнијих интегрисаних развојних окружења за рад са програмским језиком Р. Интегрисано развојно окружење отвореног кода које олакшава статистичко моделирање као и графичке могућности за Р програмски језик. Први пут када се отвори RStudio видимо почетни интерфејс, који је подељен у три дела (Слика 2.).

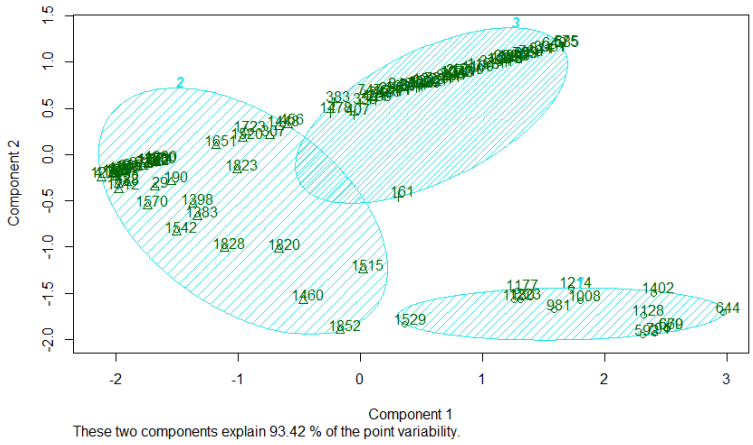


Слика 2. Почетан екран окружења софтвера RStudio

4 РЕЗУЛТАТИ И ДИСКУСИЈА

Истраживање у овом раду показује резултате анализе различитим методама кластеровања. Припремљен скуп података коришћен је за истраживање, али са различитим техникама кластеровања. Циљ истраживања био је да се учесници у саобраћају групишу према месту у којем је почињена највећа количина прекршаја.

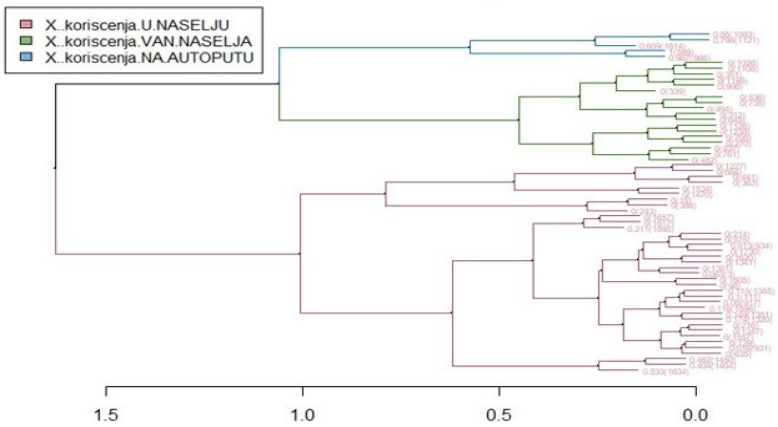
Првенствено морамо применити алгоритам учења, у нашем случају је то “K-means”. Након примене алгоритма кластери могу бити представљени исцртавањем различитих дијаграма и дендограма у “R Studio” софтверу. За почетак ћемо исцртати једноставан дендограм који приказује резултате кластеровања применом алгоритма “K-means” у дводимензионалном облику (Слика 3.)



Слика 3. Дендограм који приказује дводиментиони приказ резултата кластеровања

4.1 Хијерархијско кластеровање

У наставку је коришћено хијерархијско кластеровање дељењем због чињенице да је ефикасније од агломеративног кластеровања јер има нижу нотацију сложености O(n2). Такође, тачније је, јер агломеративно кластеровање доноси одлуке узимајући у обзир локалне обрасце без разматрања глобалне дистрибуције података. Те ране одлуке не могу да се преокрену и то утиче на резултат дат хијерархијским агломеративним кластеровањем. Може постојати неколико најближих парова подгрупа, али се бира само један пар при свакој итерацији, након чега се процес итерације понавља од почетка. Другим речима, примењена је пермутација на елементе X и поново покренут алгоритам.

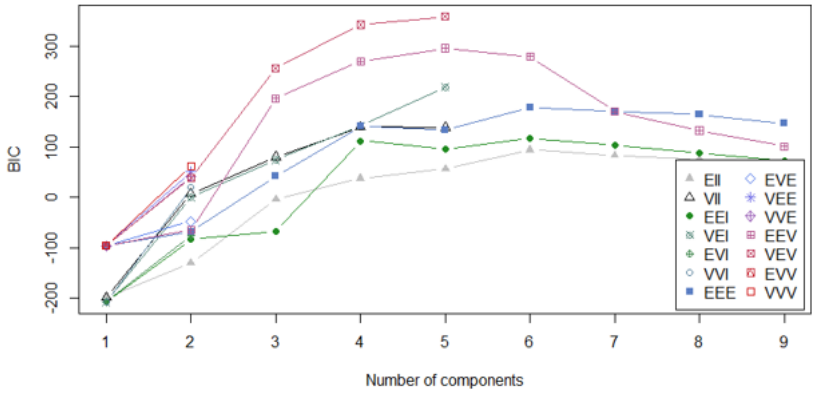


Слика 5. Дендограм приказа хијерахијског кластеровоања у облику стабла

Из дендограма на слици 5. можемо закључити да највише има прекршаја у саобраћају који се праве у насељу, док најмање прекршаја су направљени на аутопуту. Иста техника може бити представљена дендрограмом са груписањем по редовима надоле са вредностима.

4.2 Кластеровање засновано на Бајесовом информационом критеријуму (БИЦ)

Бајесови фактори, приближни Бајесовим информационим критеријумима (БИЦ), успешно су примењени на проблем утврђивања броја компоненти у моделу и за доношење одлуке које се међу три партиције најприближније подударају са подацима за дати модел.



Слика 6. График приказа Бајесовог информативног критеријума (БИЦ)

На слици 6., график приказује Бајесов информациони критеријум (БИЦ) за методе засноване на моделима примењеним на податке о саобраћајним прекршајима. Први локални максимум се јавља за неограничени модел са три кластера. Упоређени су девет различитих модела у погледу вероватноће. Најбољи модел је онај који садржи најнижу БИЦ вредност. Сваки модел приказан је помоћу 14 различитих индикатора. У нашем случају најбоље БИЦ вредности по моделима добијају се помоћу EII индикатора, комбиновањем HC и EM алгоритма над моделом λI. Осим код другог модела. Код другог модела најбоља БИЦ вредност се добија помоћу EEI индикатора применом EM алгоритма над моделом λA.

Табела 1. Легенда графика приказа Бајесовог информационог критеријума (БИЦ)

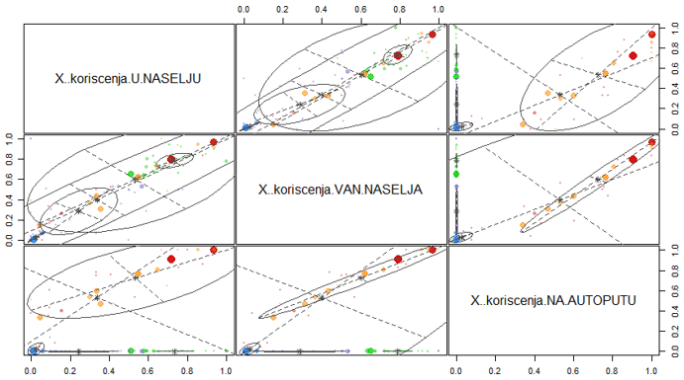
Индикатор	Модел	HC алгоритам	EM алгоритам
EII	λI	•	•
VII	λ_k I	•	•
EEI	λA		•
VEI	$\lambda_k A$		•
EVI	λA_k		•
VVI	$\lambda_k A_k$		•
EEE	$\lambda D A D^T$	•	•
EEV	$\lambda D_k A D_k^T$		•
VEV	$\lambda_k D_k A D_k^T$		•
VVV	$\lambda_k D_k A_k D_k^T$	•	•

Опис индикатора:

- EII: једнака запремина, округлог облика (сферна коваријација)
- VII: променљиве запремине, округлог облика (сферна коваријација)
- EEI: једнака запремина, једнак облик, оса паралелна оријентација (дијагонала коваријација)
- VEI: променљива запремина, једнак облик, оса паралелна оријентација (дијагонала коваријација)
- EVI: једнака запремина, различитог облика, оса паралелна оријентација (дијагонала коваријација)
- VVI: променљива јачина звука, различит облик, једнака оријентација (дијагонала коваријација)
- EEE: једнака запремина, једнак облик, једнака оријентација (елипсоидна коваријација)
- EEV: једнака запремина, једнаког облика, различите оријентације (елипсоидна коваријација)
- VEV: променљиве запремине, једнаког облика, различите оријентације (елипсоидни коваријација)
- VVV: променљива запремина, променљив облик, различите оријентације (елипсоидни коваријација)

4.3 Кластеровање на основу модела

У наставку биће приказан пресек који означава класификацију саобраћајних прекршаја, који податке прегради у три групе. Променљиве имају следећа значења: у насељу, ван насеља и на аутопутевима. Кластери се преклапају, у сферичном су облику и далеко су. Као резултат тога, многе процедуре груписања не би добро функционисале за ову апликацију.



Слика 3. Кластеровање засновано на груписању по моделу неизвесности

Слика 7. приказује пројекцију три-кластер класификације добијене методом са једном везом (најближа-комшиница), стандардним “k-means“ и методом заснованом на моделу за неограничену гаусијску мешавину за шест различитих Гаусиан модела.

5 Закључак

Резултати представљају област у којој се обавља већина саобраћајних прекршаја. То значи да посебну пажњу треба посветити тој области, у нашем случају окружењу у насељу. Сви алгоритми имају исти образац и он је повезан са израчунавањем растојања између узорака. Предност је сложеност алгоритма, који истовремено може бити неповољан ако је сложеност превелика (претпоставља се дуже извршавање израчунавања груписања).

Подаци се прикупљају са портала отворених података, тако да се такви "сирови" подаци не могу директно користити. Пре употребе, подаци морају да се обрађују техникама као што су серијализација и подешавање размере (скалирање). У сврхе истраживања у овој области пожељно је чешће покретање програма са ажурираним подацима. Посебан допринос био би отворени “API“ који би приказао изабране податке у реалном времену.

У одабраном скупу података неке променљиве су важније од других. Променљиве попут процента употребе у одређеном подручју саобраћаја важније су од возила, типа индикатора прекршаја и године прикупљања података. Резултати груписања такође су везани за подручја са већином и најмање саобраћајних прекршаја. Кључни фактор који одређује резултат је проценат употребе у одређеној области саобраћаја. На пример, без обзира на возило у којем се прекршај направио, оно би могло да буде фокусирано углавном на област где се догодио саобраћајни прекршај.

Имајући у виду добијене резултате може се извући неколико закључака:

- Анализа кластера могла би успешно да се спроведе у решавању проблема везаних за саобраћајне незгоде и све поменуте технике кластеровања имају своје предности и мане;
- Постоји значајна корелација између понашања учесника у саобраћају и коришћења сигурносних система и мобилних телефона
- Свака врста анализе кластера која се користи је значајна и комплементарна са информацијама кластера.

У пракси, резултати добијени груписањем могу се користити за одређивање нових закона о безбедности на путевима или за промену постојећих. Утврђивањем области у којој се обавља већина саобраћајних прекршаја, експерти могу да се усредсреде на законе о безбедности на путевима само за ту област. Усредсређивањем само на те области може се добити квалитетнији закон о безбедности на путевима који може да смањи број саобраћајних прекршаја.

Одрађено истраживање у овом раду отворило је могућности за даље побољшање груписањем на различите начине. Груписање би могло да се користи за одређивање области у којој се обавља већина саобраћајних прекршаја, а затим поново груписање скупа података само за ту одређену област. Предложеним побољшањем могли би да се утврде детаљи попут тога у ком возилу се обавља већина саобраћајних прекршаја и која врста саобраћајног прекршаја се углавном догађа.

Прикупљањем тих детаља стручњаци за стварање закона о безбедности на путевима не смеју да се фокусирају само на област на којој се обавља већина саобраћајних прекршаја, већ и на специфична возила и врсту прекршаја. То ће резултирати специфичнијим законима са фокусом на најпроблематичније безбедносне факторе. Новостворени закони ће сигурно смањити број саобраћајних прекршаја који су се догодили. Такође се може ставити опрема за безбедност саобраћаја како би се спречили неки од прекршаја.

Литература

[1] Закон о безбедности саобраћаја на путевима ((—The Official Gazette of the Republic of Serbia ‖ , No. 41/2009, 53/2010, 101/2011, 32/2013 – Constitutional Court decision, 55/2014, 96/2015 – other law, 9/2016 – Constitutional Court decision, 24/2018, 41/2018, 41/2018 – other law, 87/2018 and 23/2019)

[2] Стратегија развоја водног саобраћаја Републике Србије, 2015 – 2025, доступно на: http://aler.rs/files/STRATEGIJA_razvoja_vodnog_saobracaja_Republike_Srbije_od_2015_do_2020_godine_Sl_gl_Rs_br_3_2015.pdf, Последњи приступ 4. јануар 2021.

[3] Portal otvorenih podataka, https://data.gov.rs/sr/ , Последњи приступ 4. јануар 2021.

[4] A.K. Jain, M.N. Murty, P.J. Flynn, —Data clustering: A review ‖ , ACM Comput. Surveys, vol. 31, no. 3, pp. 264‑323, 1999.

[5] W. Yang, H. Long, L. Ma, H. Sun, —Research on Clustering Method Based on Weighted Distance Density and K-Means ‖ , Procedia Computer Science, vol. 166, pp. 507‑511, 2020.

[6] D. Hofmeyr, —Degrees of freedom and model selection for k-means clustering ‖ , Computational Statistics & Data Analysis, vol. 149, 2020.

[7] W. Wei, Liang. J, X. Guo, P. Song, Y. Sun, Hierarchical division clustering framework for categorical data, Neurocomputing, vol. 341, pp. 118–134, May 2019.

[8] G. Schwarz, —Estimation the Dimension of a Model", The Annals of Statistics, vol.6, pp. 461-464, 1978.

[9] B. Zhou, J. Hansen, —Unsupervised audio stream segmentation and clustering via the Bayesian information criterion ‖ , In Proceedings of the 6th International conference on spoken language processing, Bejing, China, 2000.