Математички факултет

Београд, Студентски трг 16

КЛАСТЕРОВАЊЕ МОНОНУКЛЕАРНИХ ЋЕЛИЈА ПЕРИФЕРНЕ КРВИ СЕКВЕНЦИОНИРАНИХ ТРАНСКРИПТОМИКОМ ПОЈЕДИНАЧНИХ ЋЕЛИЈА

– семинарски рад из Истраживања података 2 –

Ментори:

проф. др Ненад Митић

проф. др Владимир Брусић

Студент:

Лазар Васовић, 99/2016

Београд, јун 2020

С А Д Р Ж А Ј

Страна

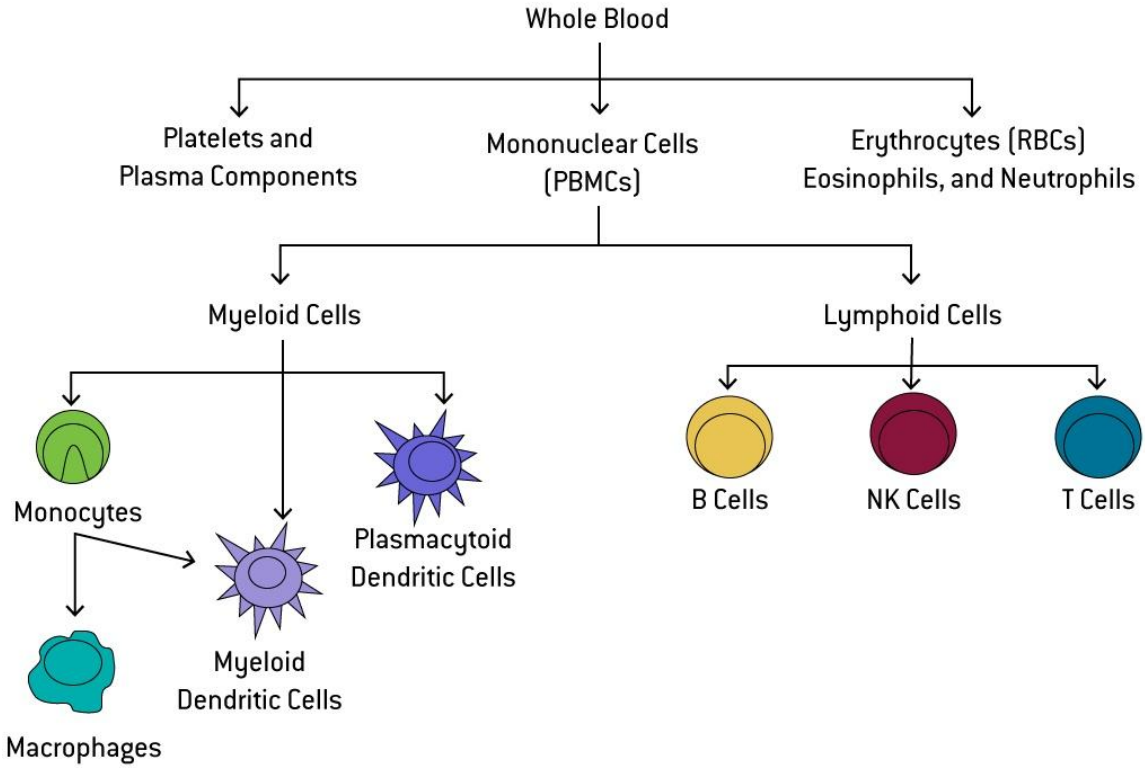
1. УВОДНА РЕЧ

1.1. Биотехнолошки оквир

Протичући циркулаторним системом човека, крв кроз тело преноси хранљиве материје и кисеоник. Већински део ове функционално најважније телесне течности чини крвна плазма, чији је главни састојак вода, за којом следе беланчевине (крвни протеини), разне соли и друге органске и неорганске материје, као и глукоза (шећер). У плазми су потопљене и **крвне ћелије**, у оквиру којих се издваја више типова, према различитим критерујимима поделе. По функцији, издвајају се црвена (еритроцити, са транспортном улогом) и бела крвна зрнца (леукоцити, са улогом у имунском систему), као и крвне плочице (тромбоцити, са коагулационом улогом, у згрушавању крви). По облику једра (централне органеле која садржи највећи део генома, уграђен у ДНК), издвајају се оне са једним лоптастим једром (**мононуклеарне ћелије**), оне без једра (ануклеарне ћелије), као и оне чије једро садржи неколико режњева и зрнца испуњена ензимима (грануларне ћелије). По месту у крвотоку, издвајају се **ћелије периферне крви**, која слободно протиче телом, док је остатак крви и у њој садржаних ћелија везан за неко одређено место или орган – лимфни систем, јетру, слезину, коштану срж...

Крвне ћелије које одликује постојање једног лоптастог једра, а које се налазе у периферној крви, називају се **мононуклеарне ћелије периферне крви** (енгл. *peripheral blood mononuclear cells*, **PBMC**). Постоји неколико различитих подгрупа оваквих ћелија, од чега се издваја пет главних – **Б ћелије** (BC), **Т ћелије** (TC), **ћелије убице** (NK), **моноцити** (MC) и **дендритске ћелије** (DC). Учесталост ових подтипова у крви знатно се разликује од особе до особе, а кроз време може варирати и на нивоу једне особе. Груба процена, до које се дошло комбинацијом претходних процена, јесте да TC чине 40–70%, MC 10–30%, BC 5–15%, NK 5–10%, а DC 1–2% људских PBMC-ја[1].

Изложени модел са пет група је сведена верзија правог модела, који је сложенији и хијерархијске природе. Прва три набројана типа чине надгрупу агранулоцита односно лимфоидних ћелија (лимфоцита), док последња два спадају у надгрупу мијелоидних ћелија. Ове две надгрупе, када се посматра претходно изнета подела крвних ћелија по функцији, заједно припадају леукоцитима (белим крвним зрнцима). Хијерархија се наставља и у другом смеру, те нпр. постоји неколико подтипова Б и Т ћелија. Остали елементи периферне крви нису мононуклеарне ћелије, тако да нису ни PBMC: еритроцити (црвена крвна зрнца) и тромбоцити (крвне плочице) уопште немају једро (ануклеари су), док другим подгрупама леукоцита (гранулоцитима – базофилима, еозинофилима и неутрофилима) оно није лоптасто, већ режњевито и са ензимским зрнцима (гранулама). Све PBMC су, дакле, **бела крвна зрнца** према функционалности, самим тим важан елемент људског имунског система, мада нису сва бела крвна зрнца PBMC. Схематски приказ описане хијерархије, уопштен тако да наглашава PBMC-је у односу на друге крвне ћелије, дат је на слици 1, која је прузета са веб-сајта швајцарске фармацеутске и биотехнолошке компаније Лонца[2].



Слика 1. Схематски приказ хијерархије ћелија у људској крви

Како спадају у крвне ћелије које чине имунски систем, PBMC имају значајну улогу у изучавању заразних болести, као и студијама у домену имунологије и аутоимунских обољења, пресађивања органа, онкологије и развоја вакцина. Користе се у проучавању рада ћелија и управљања транскрипцијом, одређивању биомаркера, као и моделовању болести. Њиховим испитивањем се може пратити здравствено стање и дијагностификовати болести крви. Лако се прикупљају, преносе и складиште, при чему постоје разрађене процедуре руковања којима се чува употребљивост узорака.

Сваки тип ћелије има својствени образац испољавања гена и изградње протеина, одређене различитим чиниоцима попут старости и фазе развоја ћелије, ткива и органа у којима се ћелија налази или тога да ли је у питању здрава ћелија или не. Најпознатији начини одређивања подскупова (типова) у оквиру PBMC-ја јесу анализа површинских рецептора техником проточне цитометрије, као и анализа њиховог транскриптомског профила (**транскриптомика**), који за сваку ћелију садржи транскриптом – број транскрипата (испољавања) сваког гена. На први начин је издвојено око стотинак различитих група, док се други, коришћен у подацима из овог рада, фокусира на пет главних. Резултат прављења транскриптомског профила сваке појединачне ћелије из узорка је врло ретка матрица података, чијих 95–99% поља су нуле. Секвенционирање **транскриптомиком појединачних ћелија** (енгл. *single-cell transcriptomics,* **SCT**) стога је прототипна технологија **великих података** (енгл. *big data*)[1]. Преко две хиљаде SCT скупова података генерисаних на овај начин од 2017. надаље јавно је доступно. Највећи део тих скупова, укључујући податке из овог рада, доступни су у бази GEO (*Gene Expression Omnibus*) америчког Националног центра за биотехнолошке информације.

Када је у питању тачан начин анализе добијеног транскриптомског профила, најзаступљенији метод, који се досад показао као успешан, јесте **кластеровање**, по могућству алгоритмима специјализованим за ретке податке великих димензија. Циљ – уочавање и прављење модела законитости у улазним подацима – постиже се тиме што се у скупу података налазе ћелије различитих типова, а на алгоритму је да открије тачну поделу на **кластере**, што су групе унутар којих су ћелије што сличније, док су ћелије из различитих кластера што различитије. То ће бити покушано и у оквиру овог рада, са нагласком на употребу неколико различитих алгоритама, као и опсежну припрему (претпроцесирање) података пре самог рада. Поред прављења претходно описаних **ненадгледаних модела**, могуће је помоћу вештачких неуронских мрежа изградити полунадгледане моделе, са бољим учинком[1], али они нису тема овог рада.

1.2. Скуп података

Обрађени скуп података чине четири узорка која се налазе у бази GEO – **GSM3330561**[3], **GSM3330562**[4], **GSM3330563**[5], **GSM3330564**[6] – а чије су главне карактеристике дате у табели која следи. Сви скупови објављени су 13. августа 2018, а ажурирани 1. септембра исте године, када је објављен рад *Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA* који им је посвећен[7].

	GSM3330561	**GSM3330562**	**GSM3330563**	**GSM3330564**
Назив	PBMC Pre	PBMC Disc Early	PBMC Disc Resp	PBMC Disc AR
Одредница	PreRx	Day+27	Day+376	Day+614
Индекс	SI-GA-A7	SI-GA-A4	SI-GA-A5	SI-GA-A6
Опис	PreRx Peripheral Blood	Early PostRx Peripheral Blood	Responding PostRx Peripheral Blood	Acquired Resistance Peripheral Blood
Димензије	17712×2082	17712×1592	17712×4684	17712×4516

Како је приказано у табели, секвенционирано је 17712 гена, док се број ћелија које чине узорке разликује – 2082, 1592, 4684, 4516. Уместо сирових формата доступних у бази GEO, коришћене су табеле задате у формату запетом раздвојених вредности (енгл. *comma-separated values,* CSV), које су ментори дистрибуирали приликом поделе тема за семинарски рад. Сви дати подаци изворно су **сирови**, што значи да поље табеле (i, j) означава колико је у ћелији *j* регистровано **испољавања (транскрипата)** гена *i*, дакле, без икакве логаритамске или друге трансформације. Како је и очекивано, све четири матрице су **ретке** – велики је удео нула поља. Ово, међутим, не значи да тих гена нема у тим ћелијама, већ само да није уочено њихово испољавање, што ове матрице чини посебно занимљивим, иако не утиче на добијене резултате.

У питању су PBMC особе која болује од ретке врсте рака коже – карцинома Меркелових ћелија. Први узорак узет је пре почетка имунотерапије (*PreRx*), а остали касније (*PostRx*), при чему је други узет након само месец дана (*Early*), трећи након отприлике једне године (*Responding*), а последњи након отприлике две (*Acquired Resistance*). Пацијент је испрва добро реаговао на терапију, али је на крају дошло до стицања отпорности на њу. Овај механизам је централна тема горепоменутог рада[7].

PBMC су из прикупљене крви изоловане уобичајеним центрифугалним поступком, а затим секвенциониране SCT методом помоћу машине Illumina HiSeq 2500. Ово је урадио тим из Центра за истраживање рака „Фред Хачинсон“ из Сијетла, Вашингтон, у својој лабораторији „Шапуи“, а под руководством онколога Кели Гарнески Полсон. Подаци су донекле претпроцесирани, а главни резултат тога је именовање гена према речнику GRCh38 (hg38) Референтног конзорцијума за геном.

Како је и карактеристично за технологије великих података, и у овом раду је примењена опсежна и пажљива додатна **припрема података**, као један од кључних корака за повећање вероватноће за добијање добрих резултата. То укључује прикладно преименовање самих података (промена назива врста и колона), евентуалне трансформације (логаритамска и остале), као и смањење димензионалности података, чиме се постижу боље перформансе алгоритама, а у неким случајевима, по принципу Окамове оштрице (бритве, жилета), добијају чак и бољи резултати са мањим трошком.

Први део тога било је читање података из дистрибуиране консултационе датотеке „common_human_list.csv“, која садржи списак **честих људских гена** и њихове номенклатуре. Уклоњене су све вредности које се не налазе у овој датотеци у пољу посвећеном GRCh38, али и записане у извештај „odbaceni.txt“. На место оних које се налазе у списку, постављени су називи настали надовезивањем одговарајућег ENSG идентификатора гена и имена према речнику Ensembl. За ове потребе је Python скриптом „json_hg38.py“ формирана мапа односно пресликавање GRCh38 у жељену номенклатуру, која је сачувана у датотеци „hg38_ensg.json“, у JSON формату. У случају да није било могуће једнозначно одредити ENSG на основу GRCh38, такви гени су одбачени и записани у извештај „dupli.txt“. Саме измене учињене су Python скриптом „priprema.py“. У исти мах је измењена и номенклатура ћелија. Оне су у сировим подацима назване шеснаестословном секвенцом нуклеотида и редним бројем датотеке, што је замењено GSM именом датотеке конкатенираним са редним бројем ћелије.

Други део припреме било је **транспоновање** добијених матрица, како би након тога колоне биле гени, а редови ћелије, пошто су оне те које се кластерују. Притом је извршено додатно **смањивање димензионалности** по обе димензије. По редовима, искључене су све ћелије у којима сумирано има мање од 1000 транскрипата (броја испољавања гена) или пак има мање од 500 ненула (испољених) гена. По колонама, искључени су гени који, када се споје све датотеке, нису испољени у бар 1% од укупног броја ћелија. Притом је проверено да све датотеке као атрибуте имају једнак број гена у истом редоследу, а у облику табеларног извештаја „GSM333056x_ispoljenost.csv“ и текстуалних „GSM333056x_nenule.txt“ и „GSM333056x_nule.txt“ издвојени су гени који се разликују између датотека. Сачувана је и матрица настала спајањем, и то у датотеци „GSM333056x.csv“. Све ово урађено је помоћу Python скрипта „spajanje.py“.

Све резултујуће датотеке са измењеним подацима назване су додавањем одговарајућег наставка на име изворног фајла. Примера ради, од првог скупа, названог „GSM3330561_PBMC_Pre.csv“, прво је настао „GSM3330561_PBMC_Pre_p.csv“ након основне припреме, а затим и „GSM3330561_PBMC_Pre_t.csv“ након додатне припреме и транспоновања. Целокупан ток припреме могуће је извршити беч скриптом

„pipeline.bat“, а у наставку следи табела са метаподацима, доступним и у датотеци „metapodaci.txt“. Број редова спојене датотеке је збир броја појединачних – 2646×7918.

	GSM3330561	GSM3330562	GSM3330563	GSM3330564
Почетно	17712×2082	17712×1592	17712×4684	17712×4516
Чести гени	16697×2082	16697×1592	16697×4684	16697×4516
Испољени	518×7918	224×7918	1122×7918	782×7918

Како је могуће приметити, димензионалност припремљених података знатно је мања од димензионалности сирових. Број гена је са 17712 сирових прво спао на 16697 честих, а затим и на двоструко мањих 7918. Број ћелија је такође опао – 518, 224, 1122, 782 – у најбољем случају, у контексту задржавања ћелија, сачувана је тек четвртина. На слици 2 је дат пример првих неколико редова и колона спојене матрице. Могуће је уочити претходно описани начин називања врста и атрибута, а већ је на овом ситном исечку приметна и ретка природа учитане матрице – од сто (20×5) приказаних поља, само су четири ненуле, при чему свака има вредност 1, мада иначе постоје и већа поља.

	A	B	C	D	E	F
1		E237491#AL669831.5	E188976#NOC2L	E188290#HES4	E187608#ISG15	E186891#TNFRSF18
2	GSM3330561_1	0	0	0	0	0
3	GSM3330561_4	0	0	0	0	0
4	GSM3330561_9	0	1	0	0	0
5	GSM3330561_12	0	0	0	1	0
6	GSM3330561_14	0	0	0	0	0
7	GSM3330561_20	0	0	0	0	0
8	GSM3330561_21	0	0	0	0	0
9	GSM3330561_23	0	1	0	0	0
10	GSM3330561_25	0	0	0	0	0
11	GSM3330561_31	0	0	0	0	0
12	GSM3330561_43	0	0	0	0	0
13	GSM3330561_44	0	0	0	0	0
14	GSM3330561_45	0	0	0	0	0
15	GSM3330561_46	0	0	0	0	0
16	GSM3330561_58	0	0	0	0	0
17	GSM3330561_61	0	0	0	0	0
18	GSM3330561_62	0	0	0	0	0
19	GSM3330561_64	0	0	0	0	0
20	GSM3330561_66	0	0	0	0	0
21	GSM3330561_79	0	0	1	0	0

Слика 2. Исечак почетног дела матрице обрађиваних података

Потенцијална замерка у вези са изложеним начином претпроцесирања могла би бити упућена на чување неколико нула колона по припремљеној датотеци. У питању су гени чије испољавање није примећено у неким узорцима крви, док у другим јесте, и то у довољној мери да покрије 1% укупног броја ћелија у све четири матрице односно спојеном узорку, над којим се удео рачуна. Ова одлука, која наизглед беспотребно повећава димензионалност без неког очигледног утицаја на формирање ненадгледаних модела, оправдава се тиме што се чувањем истих атрибута у свим датотекама олакшава упоређивање резултата тј. добијених модела. Заједничком обрадом скупови се чине **стандардизованим** и униформним у оквиру проучаване групе, што је пожељно[1].

2. НЕНАДГЛЕДАНИ МОДЕЛИ

2.1. Уводна разматрања

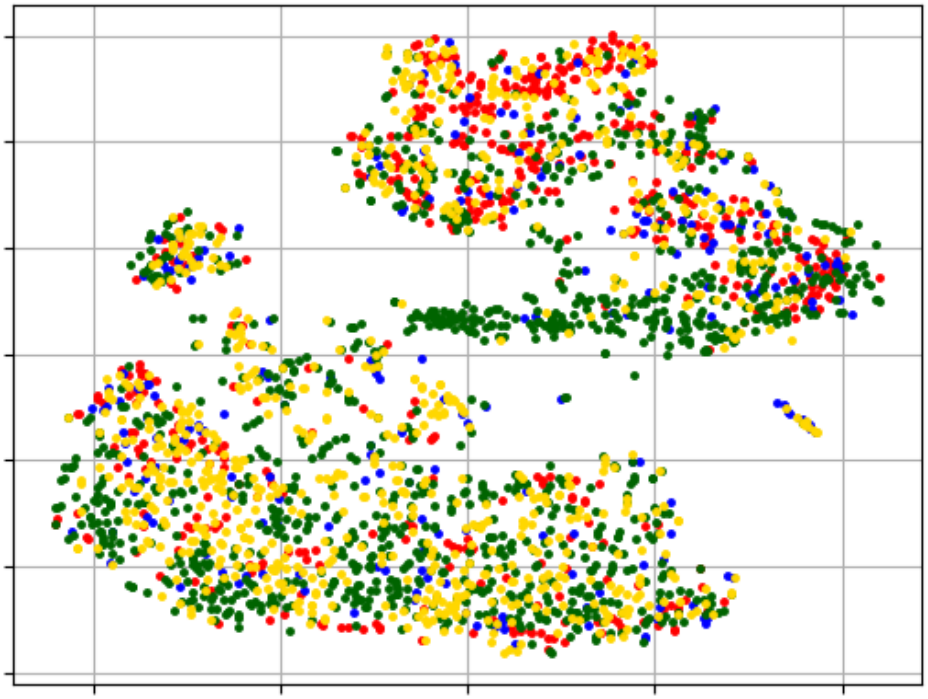
Када се говори о **ненадгледаном машинском учењу**, централна тема је потрага за неоткривеним обрасцима у неозначеном скупу података и са минималним људским надзором, који се углавном своди на подешавање параметара примењеног алгоритма. Алгоритми овог типа учења способни су да самостално уоче неке правилности, сличности и разлике у улазу и тако направе модел законитости који одговара посматраним подацима. Та њихова особина назива се **самоорганизација**.

Најзаступљениији метод анализе транскриптомског профила PBMC-ја, који се досад показао као задовољавајуће успешан, јесте кластеровање, управо пример ненадгледаног учења. **Кластер анализа** у задатом случају представља проналажење група ћелија таквих да су ћелије у групи међусобно сличне (блиске, повезане), док су ћелије у различитим групама међусобно различите (удаљене, неповезане). Растојања унутар група, које се називају кластерима, треба, дакле, да буду минимизована, док су растојања међу кластерима максимизована. Тачно откривање поделе на кластере је на самом алгоритму, уз евентуалну спецификацију и каснију измену параметара.

Кластеровање је метод који је примењен и у оквиру овог рада, са нагласком на употребу **неколико различитих алгоритама**. Коришћене технологије су програмски језик Python са модулима за машинско учење *sklearn*, *pyclustering* и *MiniSom*, као и специјализована апликација за кластеровање (g)CLUTO[8]. Примењене су како технике партиционог односно деобног, тако и хијерархијског кластеровања. Сви употребљени алгоритми, не рачунајући евентуално проглашавање неке ћелије за аномалију, примењују ексклузивно односно искључиво (свака ћелија припада тачно једном кластеру), нефази односно нерасплинуто (свака ћелија припада тачно једном кластеру са тежином 1, а свим осталима са тежином 0, дакле не припада им ни у којој мери), комплетно односно потпуно (свака ћелија припада неком кластеру, изузетно ниједном ако се прогласи за аномалију), хетерогено односно разнородно (не форсира се једнакост величина свих кластера, те су неки и знатно већи или мањи од других, што је природно с обзиром на прилично неравномерну расподелу PBMC-ја) кластеровање. Кластери су, у зависности од алгоритма, засновани на центру, суседству или на густини, при чему се за рачунање растојања између ћелија користе различите метрике (мере) простора. Квалитет резултата исказан је описном оценом која говори о уклапању модела у очекивање, док су они који највише обећавају проверени уобичајеним унутрашњим критеријумом провере – сенка коефицијентом, који бројчано исказује однос повезаности (кохезије) унутар кластера и раздвојености од ћелија изван њих.

Уочена је и разлика у циљу кластеровања појединачних скупова и спојеног, при чему је код првих фокус на одређивању типа ћелије, а код другог и на одређивању скупа из ког је ћелија потекла, односно тога из које је фазе терапије. За потребе другог задатка у скрипту „ispitivanje.py“ оцењен је теоретски савршен модел, у коме постоје четири кластера, при чему први чине тачно ћелије из датотеке GSM3330561 и тако даље. Добијени резултати су у неку руку изненађујући, пошто је, независно од узете метрике, сенка коефицијент овакве поделе негативан, што указује на врло лош квалитет кластеровања. Овакво стање ствари најављује да је у питању тежак проблем, те да су ћелије прилично нераздвојиве према фази терапије. Ово је потврђено приликом рада, пошто ниједан добијени модел ни у којој мери не подсећа на поделу по датотекама пре спајања. Сваки налази кластере према некој мери растојања, а оне су у конкретном случају бесмислене, пошто се групе ћелија очигледно изразито преклапају у простору.

Успут је, за потребе визуелизације, као и ради касније поновне употребљивости, одређена и **t-SNE** (енгл. *t-distributed stochastic neighbor embedding*) трансформација односно димензиона редукција улазних скупова на две димензије, а резултати сачувани у одговарајућим „joblib“ датотекама. Графички приказ већ поменутог идеалног модела – тачне поделе по датотекама – сачуван је као „GSM333056x.png“, а приложен кроз слику 3 из наставка. Ћелије су представљене тачкама у редукованој равни, док боје сведоче о припадности одговарајућем кластеру тј. управо фајлу у случају изложеног идеалног модела. И у наставку рада, бојама је исказана припадност, и то црвена означава први кластер (ознака 0 у датотекама са списком ознака), зелена други (ознака 1), плава трећи (ознака 2), жута четврти (ознака 3), а љубичаста пети (ознака 4). Због природе проблема (пет типова ћелија), нису разматрани модели са више кластера. Како је приметно и на слици, баш свака боја налази се у баш сваком делу простора, без јасне раздвојености, мада се уочавају области у којима доминирају одређене групе.



Слика 3. Идеални модел кластеровања спојене датотеке – црвени кластер су ћелије преузете из првог фајла, зелени из другог, плави из трећег, жути из четвртог

2.2. Репрезентативне тачке

На почетку, размотрено је кластеровање засновано на **репрезентативним представницима**. Кластер је код њих скуп објеката таквих да је сваки објекат из кластера ближи прототипу (центру) тог кластера у односу на прототипе осталих кластера. Центар кластера може бити центроид (просек свих тачака у кластеру) или медоид (најрепрезентативнија тачка у кластеру). Задаје се одређени број репрезентативних представника (тиме и број кластера), након чега се итеративно миминизује растојање припадајућих елемената од центра кластера којем припадају.

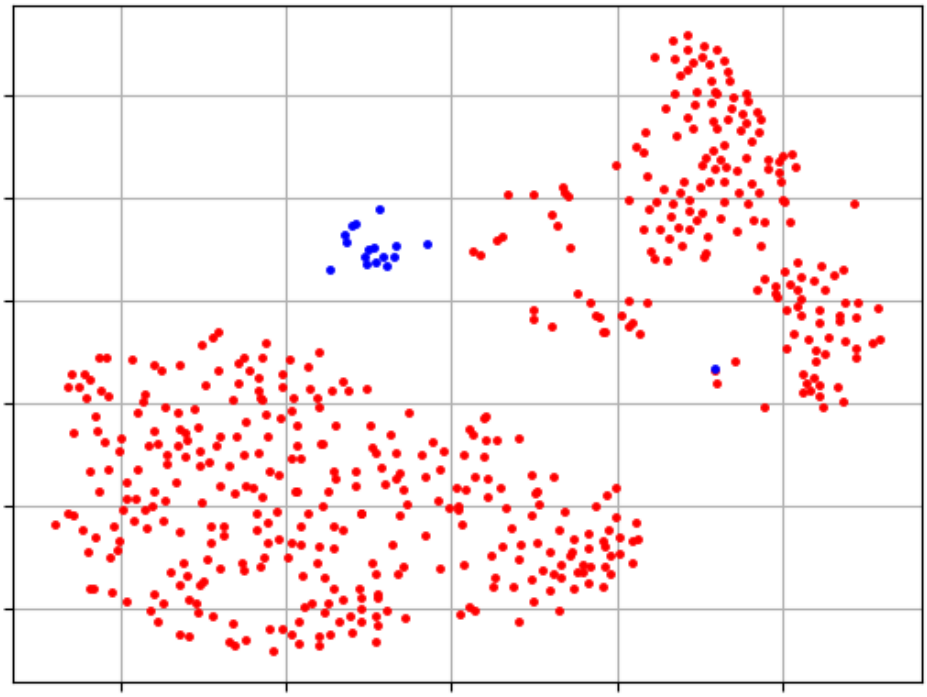
Најзначајнији алгоритам овог приступа јесте алгоритам ***k*-средина**. Код њега су прототипи центроиди, док се као приступ одређивању кластера користи партиционо кластеровање. Број кластера је *k* и он је, како је већ наглашено, улазни параметар. На почетку се свака тачка додељује кластеру са најближим центроидом. Затим се у

итерацијама израчунавају нови центроиди, као и припадности тачака кластерима. Разматра се нпр. еуклидско растојање до најближег репрезентативног представника.

Остали важни представници су алгоритми ***k*-медијана** и ***k*-медоида**, који се од претходног разликују по избору прототипа, овде углавном медоида. Код првог се уместо аритметичке средине (просека) тачака узима њихова медијана, док се код другог разматрају искључиво тачке из иницијалног скупа. Оба приступа смањују осетљивост на елементе ван граница, а главна мера је Менхетн растојање уместо еуклидског.

Када су у питању скупови из рада, скриптом „kmeans.py“ формирано је чак 180 различитих модела заснованих на репрезентативним представницима. Иницијални представници одабрани су хеуристичком методом *KMeans++* у случају прва два алгоритма односно насумично у случају трећег. За сваки скуп је тражено између два и пет кластера, при чему су коришћене три мере растојања између инстанци – Менхетн, еуклидско и косинусно. Називи модела одређени су по шаблону именовања типа „<datoteka>_<k><algoritam>_<mera>.joblib“. Графички приказ сваког сачуван је у PNG формату. У датотеци „ksilhouette.txt“ сачувани су и сенка коефицијенти модела, као и мапа расподеле кластера за сваки, у којој се ознака кластера слика у његову величину.

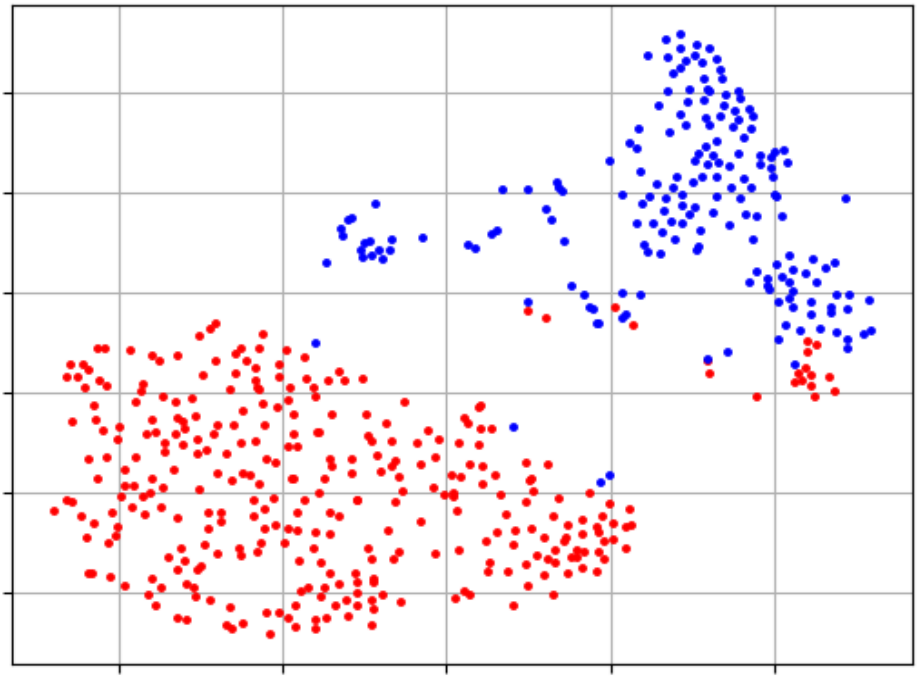
Добијени резултати се у великој мери међусобно разликују, а они који се чине најбољима издвојени су у наставку. Када је и питању прва датотека, највећи сенка коефицијент од 0,47 има модел 2-средина са Менхетн растојањем (слика „GSM3330561_2means_man.png“), приказан на слици 4. Ипак, иако је приметно да је раздвајање плавих тачака релативно добро, превелик је удео разбацаних црвених.



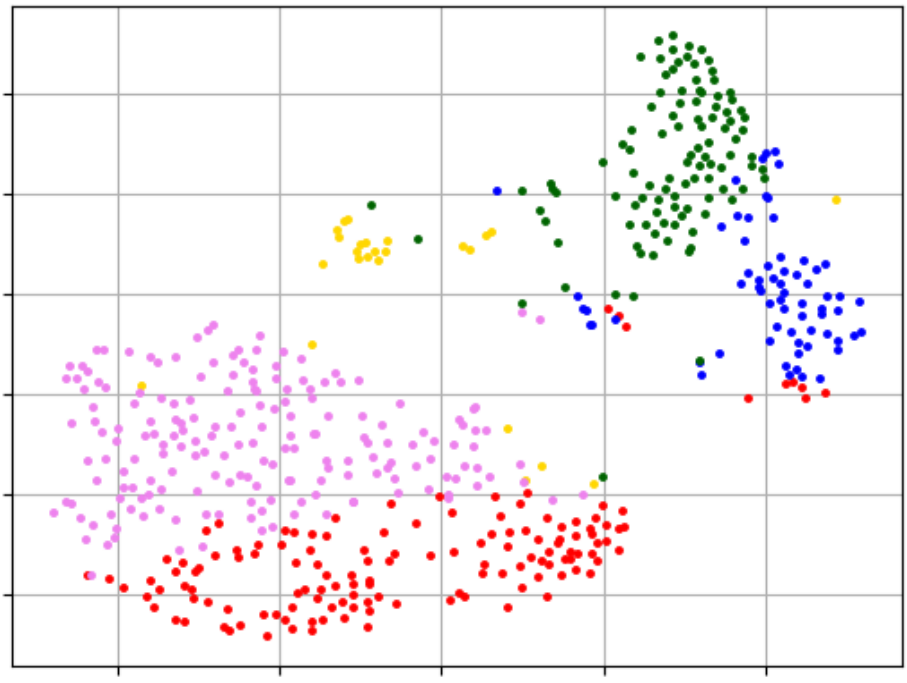
Слика 4. Модел 2-средина са Менхетн растојањем за прву датотеку, у сачуваним сликама назван „GSM3330561_2means_man.png“

Насупрот томе, најравномернију расподелу ћелија по кластерима има модел 2-средина са косинусним растојањем („GSM3330561_2means_cos.png“) и сенком од 0,4. И његова визуелизација дата је у наставку, на слици 5. Остали модели, са више кластера и другим различитим параметрима, дали су лошије резултате, што је вероватно последица чињенице да се и са самих слика, и овог пута направљених дводимензионом

t-SNE редукцијом, чини да стварно постоје два кластера када се сагледа положај у простору. И визуелним упоређивањем са осталим моделима може се доћи до закључка да се косинусни 2-средина модел најбоље показао. Други са два кластера не деле тачке тако добро, док се за оне са више чини да без очигледног правила деле тачке у групе.



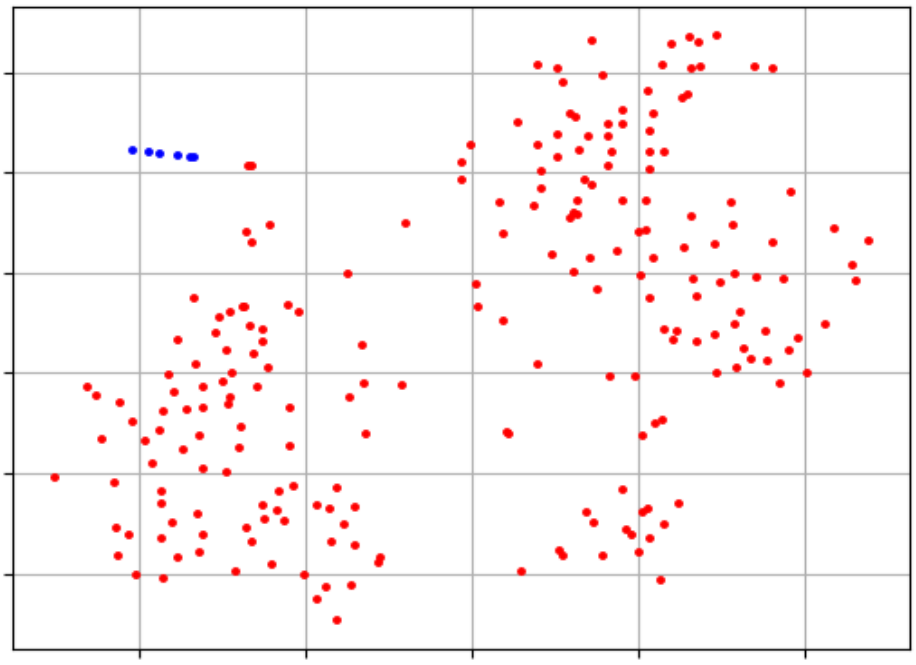
Слика 5. Модел 2-средина са косинусним растојањем за прву датотеку, у сачуваним сликама назван „GSM3330561_2means_cos.png“



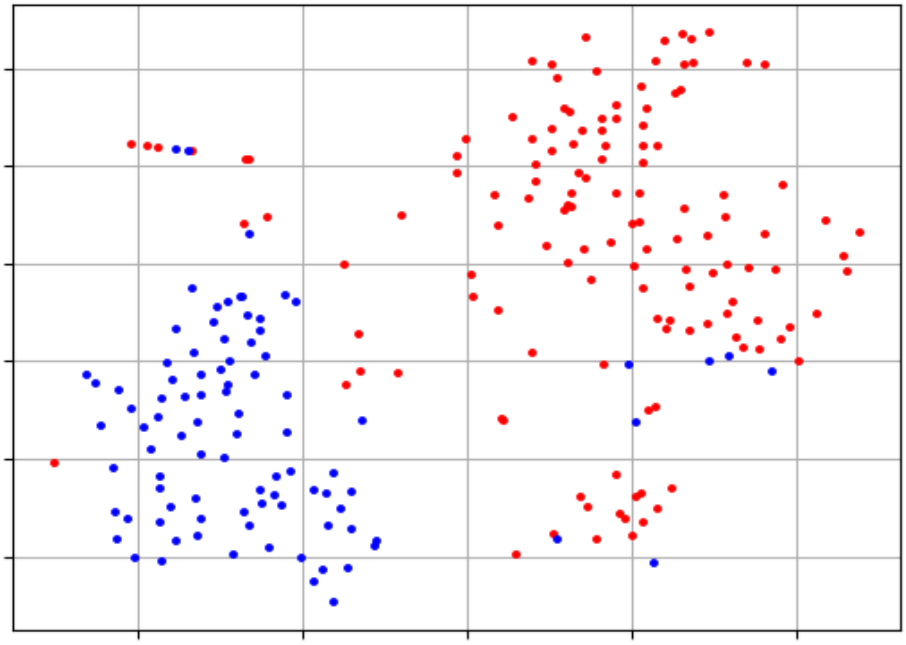
Слика 6. Модел 5-медијана са косинусним растојањем за прву датотеку, у сачуваним сликама назван „GSM3330561_5medians_cos.png“

Посебна пажња посвећена је моделима са пет кластера, пошто би то потенцијално могла бити подела на пет типова PBMC ћелија. У том контексту, колико-толико добар сенка коефицијент од 0,27 по вредности има косинусни 5-медијана модел („GSM3330561_5medians_cos.png“), приказан на слици 6. Расподела ћелија према кластерима је 27% : 13% : 20% : 5% : 36%, што у некој мери одговара процењеној расподели PBMC-ја у људској крви, изложеној у уводу рада, мада са већом активношћу иначе ретких ћелија, што може бити последица чињенице да је узорак болесне особе.

Када је у питању друга датотека, највећи сенка коефицијент од чак 0,86 има модел 2-средина са еуклидским растојањем (слика „GSM3330562_2means_euc.png“), али са истим проблемом прилично неравномерне расподеле као и код прве датотеке. Најравномернију расподелу по кластерима има косинусни 2-медоида модел (слика „GSM3330562_2medoids_cos.png“) са сенком 0,27. Оба су приказана у наставку, као слике 7 и 8. Гледајући моделе са пет кластера, опет се најбоље показао косинусни 5-медијана модел, али то најбоље је са три занемарљиво мала кластера (1-2% удела) и само два озбиљна, тако да није визуелизован, али би се потенцијално могао протумачити као нагли пад активности ретких (малигних) ћелија на почетку терапије.

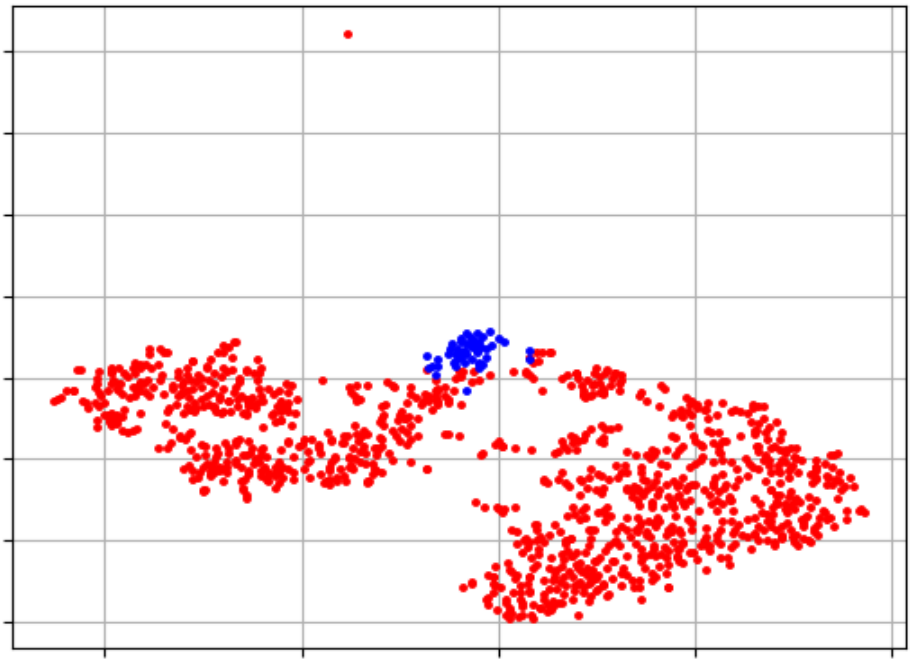


Слика 7. Модел 2-средина са еуклидским растојањем за другу датотеку, у сачуваним сликама назван „GSM3330562_2means_euc.png“

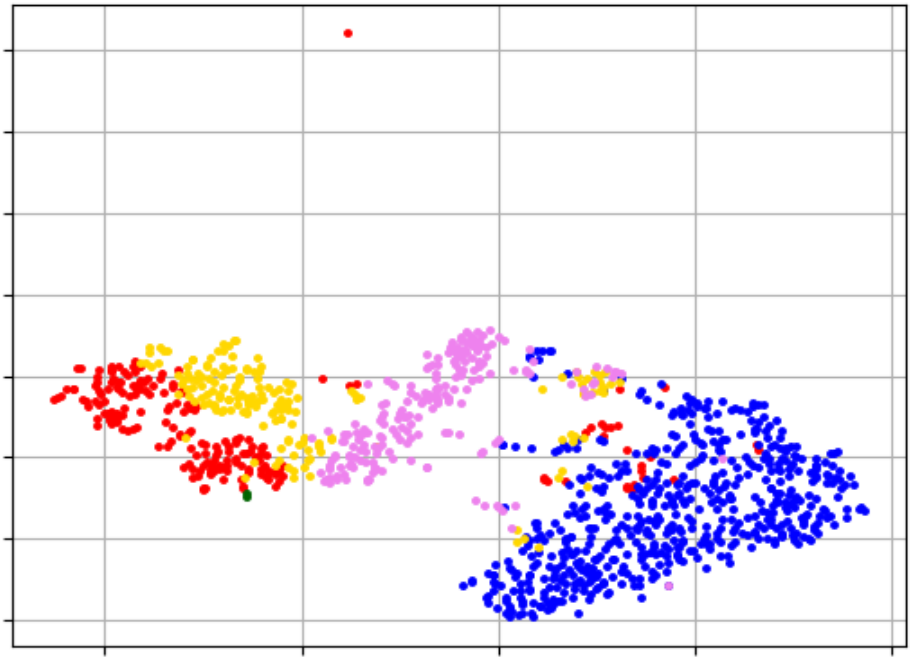


Слика 8. Модел 2-медоида са косинусним растојањем за другу датотеку, у сачуваним сликама назван „GSM3330562_2medoids_cos.png“

Када је у питању трећа датотека, највећи сенка коефицијент од 0,58 има модел 2-медијана са Менхетн растојањем (слика „GSM3330563_2medians_man.png“) са слике 9, али опет са проблемом неравномерне расподеле. Многи модели прилично су равномерни, па нису посебно разматрани, а као можда најбољи издваја се косинусни 5-медијана модел (слика „GSM3330563_5medians_cos.png“) са слике 10, не толико због сенке 0,23 колико због расподеле по кластерима 17% : 50% : 1% : 15% : 17%, која доста подсећа на процењену расподелу PBMC-ја. Ово је потенцијално посебно занимљиво када се узме у обзир да је ово узорак на који терапија има највећи утицај, а резултати истог типа модела који је пре лечења указао на повећан удео посебно активних ћелија, као и нагли пад на почетку терапије, сада указују на нормализацију стања.



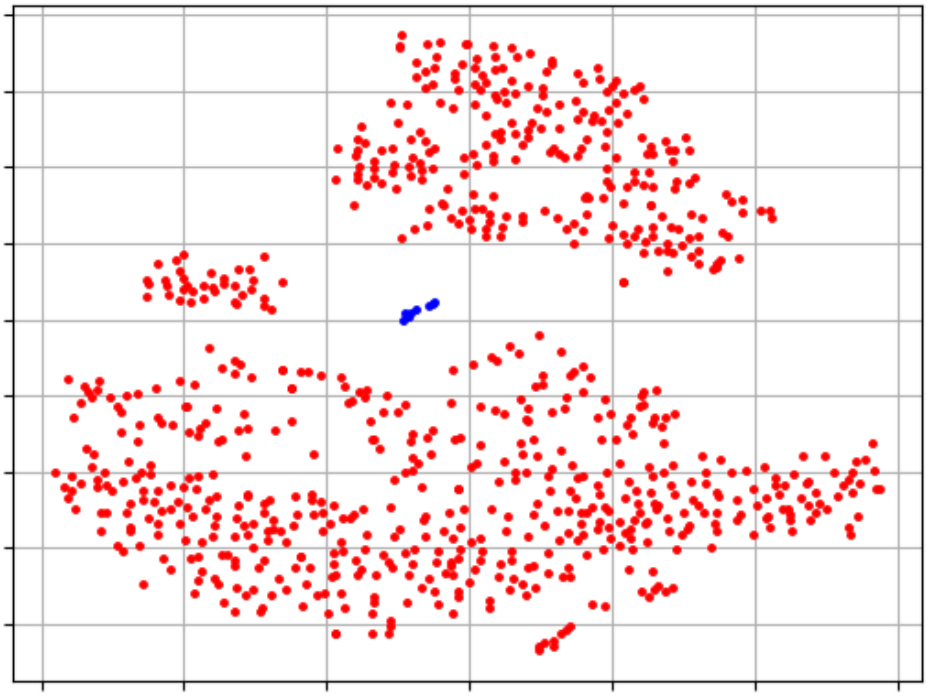
Слика 9. Модел 2-медијана са Менхетн растојањем за трећу датотеку, у сачуваним сликама назван „GSM3330563_2medians_man.png“



Слика 10. Модел 5-медијана са косинусним растојањем за трећу датотеку, у сачуваним сликама назван „GSM3330563_5medians_cos.png“

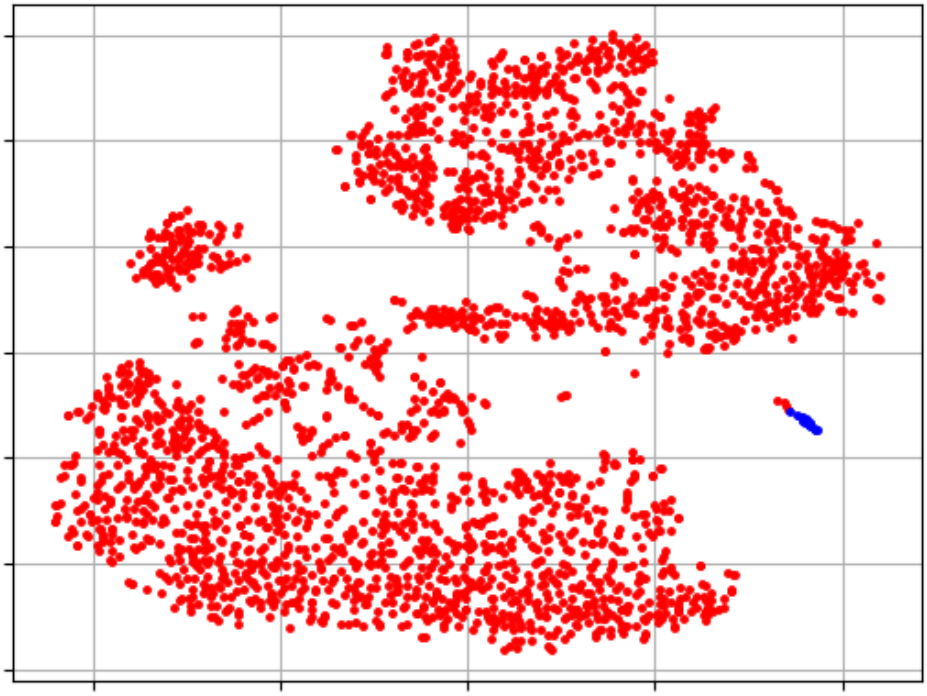
Када је у питању четврта датотека, највећи сенка коефицијент од 0,89 има модел 2-медијана са еуклидским растојањем (слика „GSM3330564_2medians_euc.png“) са слике 11, али и истим проблемом неравномерности као досад. Не постоје претерано добри равномерни модели, а на овом скупу нису неки успех постигли ни они са пет кластера – ниједан не погађа процењени удео ћелија, а углавном су и лоши сенка коефицијенти. Како је ово скуп ћелија извађених након стицања отпорности на терапију, могуће тумачење најбољег модела могло би бити да су згуснуте ћелије

препознате као мали кластер једине преостале неактивне, док промена нивоа активности праћена стицањем отпорности можда објашњава чињеницу да се више не погађа очекивани удео, чак ни код обећавајућег косинусног 5-медијана модела.

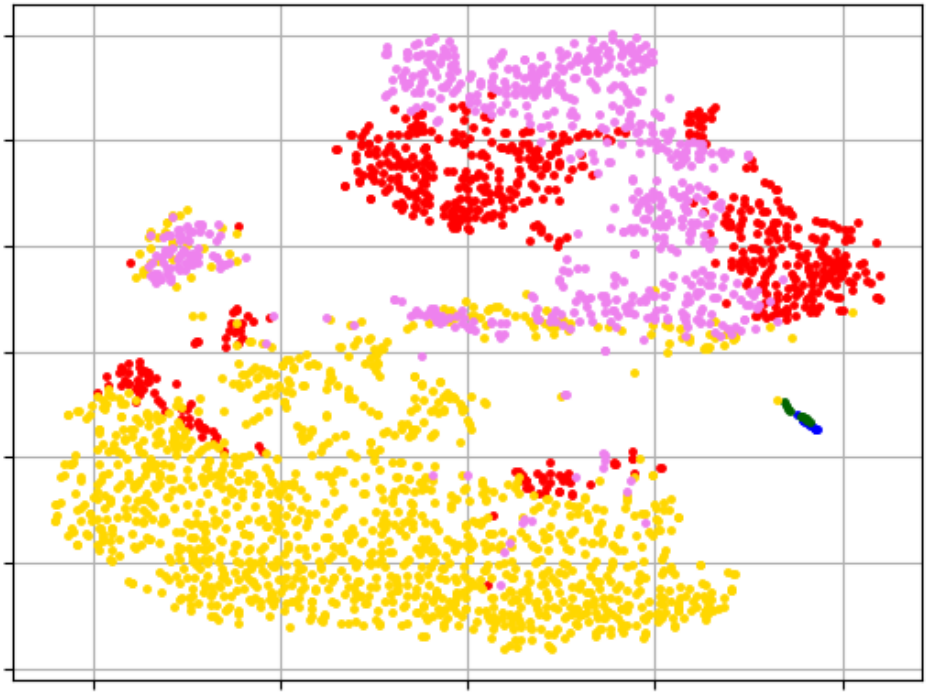


Слика 11. Модел 2-медијана са еуклидским растојањем за четврту датотеку, у сачуваним сликама назван „GSM3330564_2medians_euc.png“

Напослетку, код спојене датотеке, највећи сенка коефицијент од 0,89 има модел 2-медијана са еуклидским растојањем (слика „GSM333056x_2medians_euc.png“) са слике 12, али са истим проблемом неравномерности као досад. Ни овде нема добрих равномерних модела, док се као потенцијално добар са пет кластера још једном издваја косинусни 5-медијана модел (слика „GSM333056x_5medians_cos.png“) са слике 13. Иако му је сенка тек 0,23 по вредности, а слика показује преклапања, расподела по кластерима 23% : 1% : 1% : 51% : 24% највише личи на процењену. Ниједан модел са четири кластера на показује сличност са расподелом по почетним фајловима.



Слика 12. Модел 2-медијана са еуклидским растојањем за спојену датотеку, у сачуваним сликама назван „GSM333056x_2medians_euc.png“



Слика 13. Модел 5-медијана са косинусним растојањем за спојену датотеку, у сачуваним сликама назван „GSM333056x_5medians_cos.png“

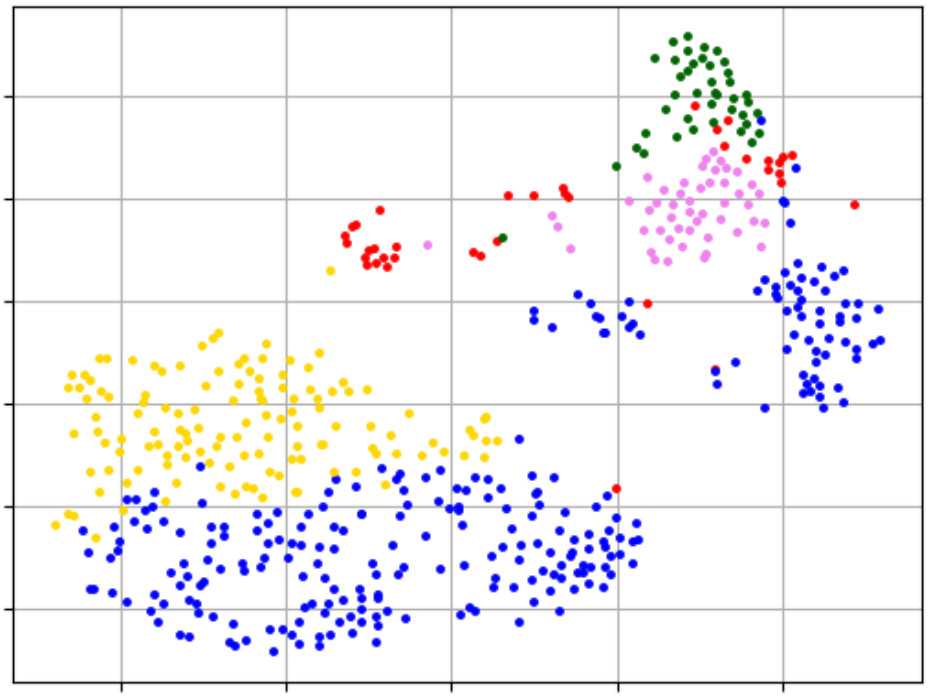
2.3. Хијерархијски модели

Следећа разматрана метода било је сакупљајуће **хијерархијско кластеровање**. Код овог модела кластери могу да садрже угнежђене друге кластере, па један елемент може да буде део више кластера на различитим нивоима хијерархије. Скуп кластера је

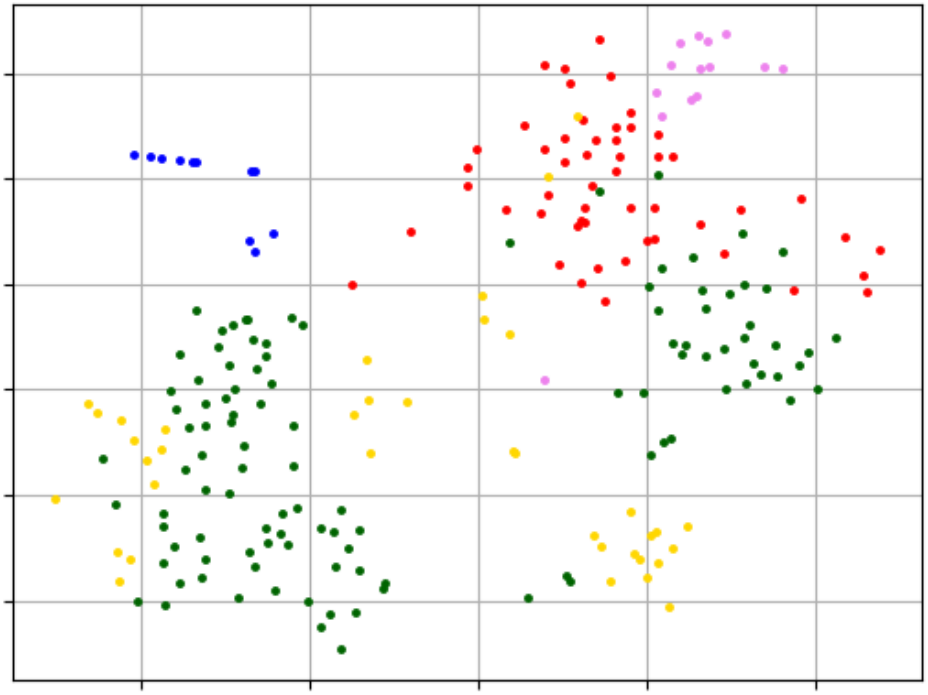
организован у облику дрвета, при чему листови садрже инстанце које се кластерују, док је корен дрвета кластер који садржи све елементе, дакле комплетан улазни скуп. Недвосмислени резултати попут оних добијених деобним кластеровањем овде се добијају пресецањем дрвета на жељеном нивоу хијерархије (нпр. најбољи број кластера) односно узимањем тог нивоа као коначног и јединог релевантног.

Оно што издваја **сакупљајуће** (*агломеративно*) хијерархијско кластеровање од других јесте смер у ком се гради хијерархија. На почетку је сваки елемент кластер за себе, а затим се постепено кластери спајају, док се не дође до једног свеобухватног супркластера. Супротан смер је одлика раздвајајућег (*дивизивног*) метода, који ће тек касније бити разматран. Битна одлика сакупљајућих модела јесте начин на који се рачуна сличност кластера, посебно након спајања. Најпознатије су минимум (најбоља, најкраћа, појединачна, *single* веза), максимум (најгора, најдужа, потпуна, *complete* веза), просек (средња, *average* веза) и Вардов метод (*Ward*, минимизација варијанси), са различитим предностима и манама, и све оне су коришћене у наставку. И овога пута је тражено два до пет кластера, уз три познате мере растојања – Менхетн, еуклидско и косинусно – а за Варда искључиво еуклидско, пошто минимизује квадратну грешку.

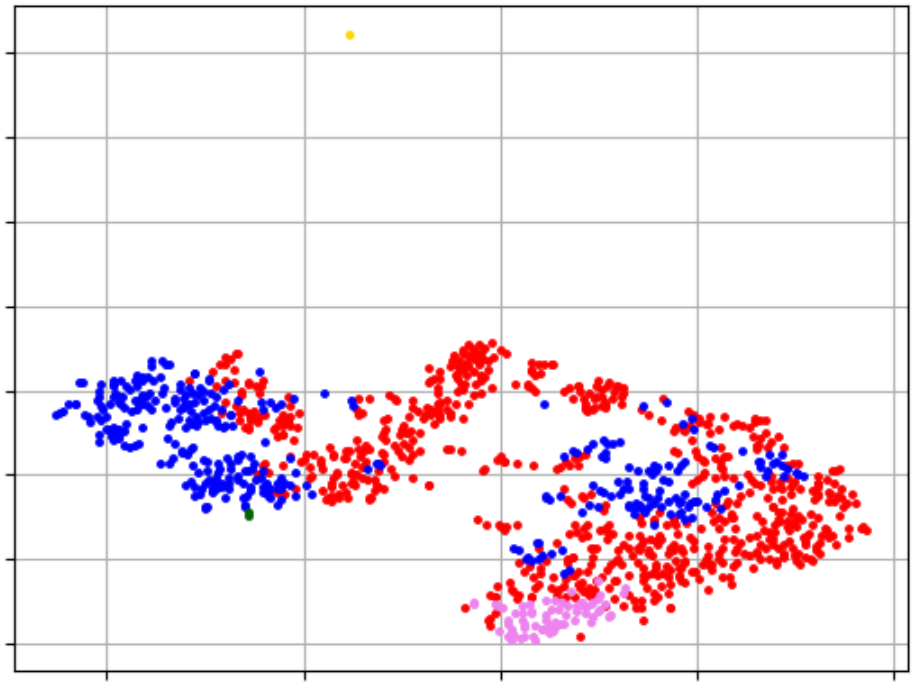
Скриптом „hijerarhija.py“ формирано је 200 различитих хијерархијских модела, за сваку могућу комбинацију параметара. Називи модела одређени су по шаблону именовања типа „<datoteka>_<k><veza>_<mera>.joblib“. Сваки је визуелизован. У датотеци „agglosilhouette.txt“ сачувани су и сенка коефицијенти модела, као и мапа расподеле кластера за сваки, у којој се ознака кластера слика у његову величину. Још једном су се као најбољи по сенци (и слични ранијим резултатима) показали модели са два кластера, посебно еуклидски са комплетном везом, али остао је проблем неравномерне расподеле. Најравномерније резултате дао је Вардов метод, са очекиваним глобуларним кластерима, док је најсличнији претходном наизглед успешном косинусном 5-медијана моделу био косинусни са пет кластера и комплетном везом. Он је и приказан за све разматране датотеке, кроз слике 14, 15, 16, 17, 18.



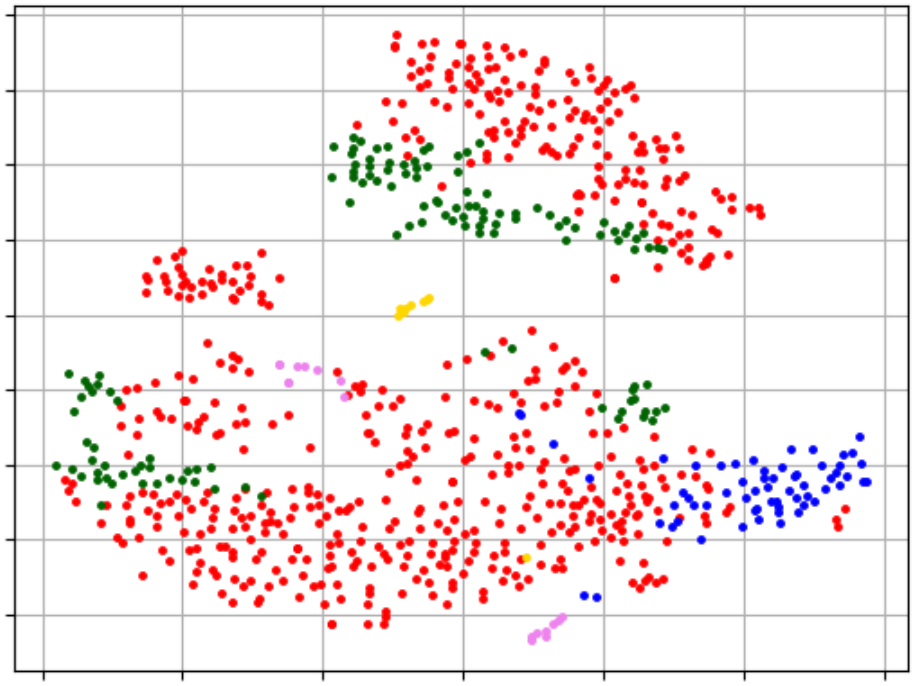
Слика 14. Сакупљајући 5-комплетни модел са косинусним растојањем за прву датотеку, у сачуваним сликама назван „GSM3330561_5complete_cos.png“



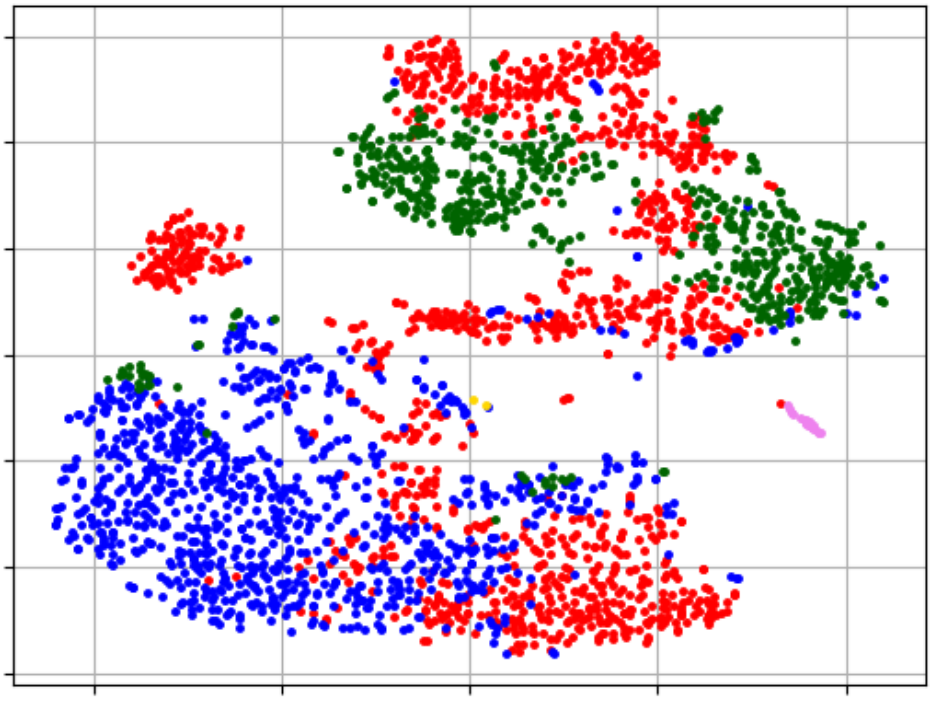
Слика 15. Сакупљајући 5-комплетни модел са косинусним растојањем за другу датотеку, у сачуваним сликама назван „GSM3330562_5complete_cos.png“



Слика 16. Сакупљајући 5-комплетни модел са косинусним растојањем за трећу датотеку, у сачуваним сликама назван „GSM3330563_5complete_cos.png“



Слика 17. Сакупљајући 5-комплетни модел са косинусним растојањем за четврту датотеку, у сачуваним сликама назван „GSM3330564_5complete_cos.png“



Слика 18. Сакупљајући 5-комплетни модел са косинусним растојањем за спојену датотеку, у сачуваним сликама назван „GSM333056x_5complete_cos.png“

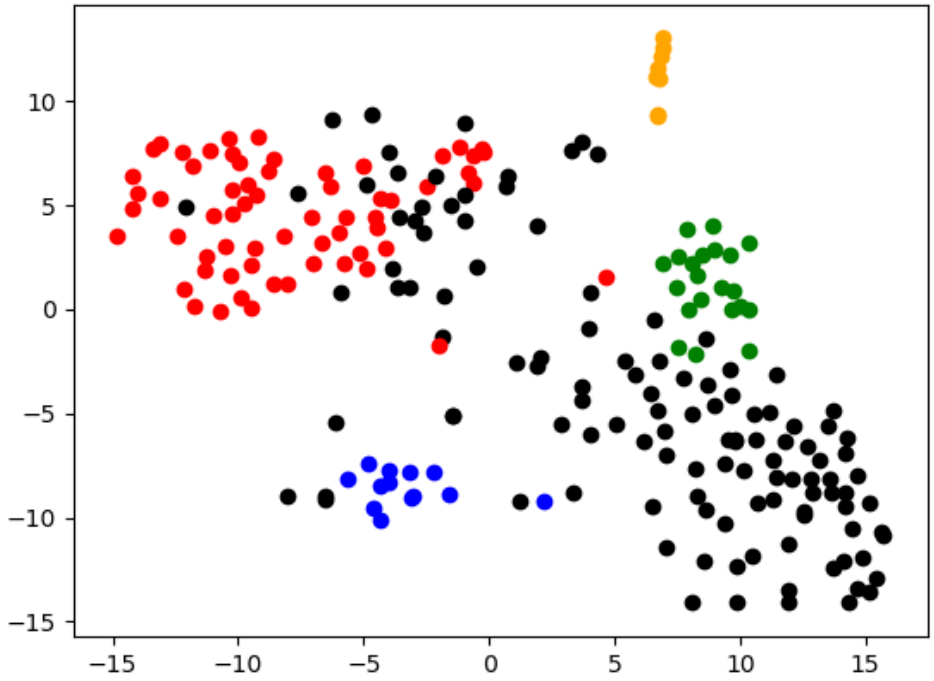
2.4. Анализа густине

Следеће разматране методе кластеровања биле су оне засноване на **густини**. Код њих кластери представљају области са великом густином тачака које су раздвојене областима са малом густином тачака. Ова карактеристика помаже у издвајању неправилних или испреплетаних кластера, као и у случајевима када су присутни шум или елементи ван граница, али је проблематична за кластере различитих густина.

Најпознатији представник ове групе јесте **DBSCAN** (енгл. *density-based spatial clustering of applications with noise*), који улазне тачке на основу параметра минималне удаљености дели у три групе – оне које припадају језгру, оне на граници и оне које чине шум (аномалију, не припадају ниједном кластеру). Сличан њему је алгоритам **OPTICS** (енгл. *ordering points to identify the clustering structure*), који гради хијерархију елемената, при чему донекле решава проблем кластера различитих густина. Осим тога, у стању је да самостално одреди најбољу вредности параметра минималне удаљености. У овом раду је примењена нешто измењена верзија доступна у модулу *sklearn*.

Када су у питању скупови из рада, ови алгоритми нису дали задовољавајуће резултате. Наиме, оба приступа при подразумеваним вредностима параметара већину инстанци проглашавају за аномалије, а остале стављају у један велики кластер, уз евентуална два или три додатна са једноцифреним бројем инстанци. Повећавањем минималног броја елемената у кластеру додатни мали нестају и све инстанце се смештају у један. Исто се дешава повећавањем минималне удаљености, док смањење тог параметра доводи до тога да свака инстанца буде проглашена елементом ван граница. Мера не утиче превише на решење, али се уочава да се косинусно растојање боље понаша од Менхетн или еуклидског, који независно од других параметара настоје да све инстанце прогласе за један кластер, те не формирају ни друге мале групе. Код OPTICS-а је осетно боље када се као начин извлачења кластера из направљене хијерархије користи специјализовани χ-метод уместо класичног DBSCAN-овског.

Скриптом „optics.py“ формирани су OPTICS модели по већ познатом шаблону именовања типа „<datoteka>_optics_<mera>.joblib“. Издвојен је и графички приказ јединог иоле доброг добијеног модела, што је DBSCAN за минималну удаљеност 0,095 и косинусно растојање за другу појединачну датотеку. Слика је пригодно названа „GSM3330562_dbscan_cos.png“ и приказана као слика 19. Као и досад, за визуелизацију је послужила дводимензиона редукција добијена алгоритмом t-SNE. Наравно, иако најбољи од добијених, модел није добар пошто одбацује пола ћелија (црна боја), али ваља напоменути да сенка коефицијент преостала четири кластера (црвена, зелена, плава и наранџаста боја) износи око једну трећину у плусу, што и није толико лоше. И са саме слике приметно је да су нецрне тачке заиста добро раздвојене, без преклапања. Урпкос томе, изостаје могућност икаквог корисног тумачења добијеног резултата, осим тога да густина изгледа није важан фактор за успешно кластеровање РВМС-ја.



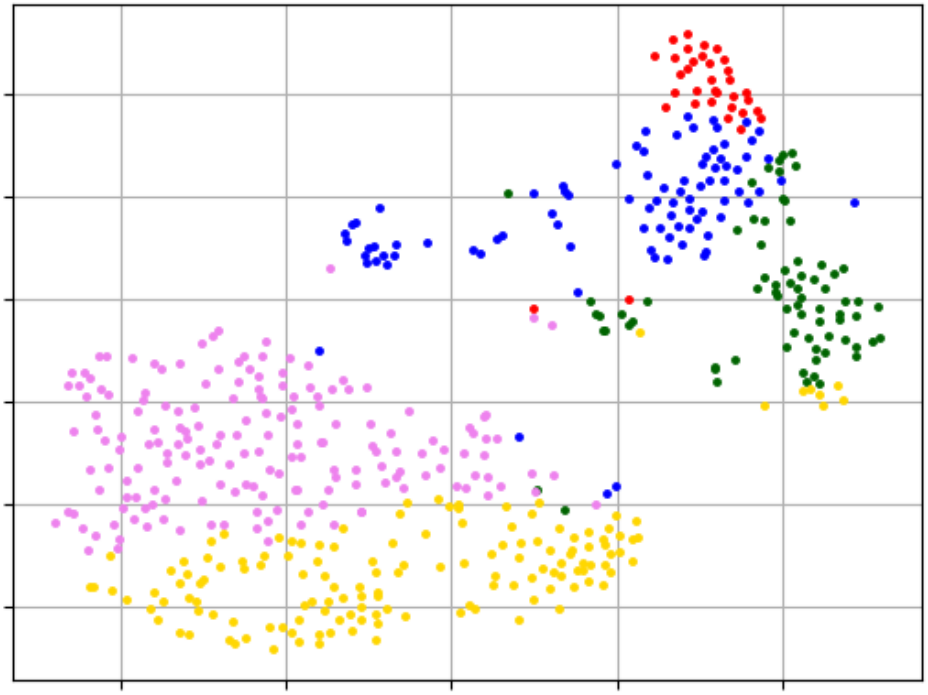
Слика 19. DBSCAN модел са косинусним растојањем за другу датотеку, у сачуваним сликама назван „GSM3330562_dbscan_cos.png“; специјално, уз четири кластера обојена уобичајеним бојама црвеном, зеленом, плавом и жутом, издвајају се црне аномалије

2.5. Самоорганизујуће мапе

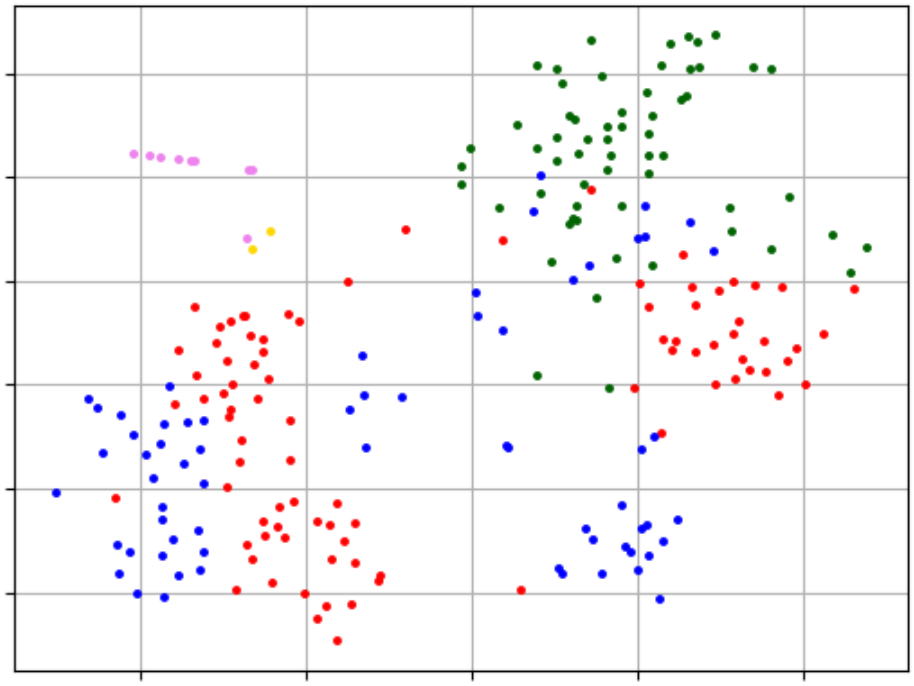
Иако је напоменуто да полунадгледани модели, који комбинују кластеровање са вештачким неуронским мрежама, нису тема овог рада, размотрен је томе сличан метод који кластере проналази својеврсним такмичењем неурона за (при)добијање инстанци. У питању су тзв. **асоцијативне неуронске мреже** (*асоцијативна меморија*), које праве асоцијације између шаблона. Обично су двослојне, а асоцијативну моћ им пружа тополошка организација неурона. Одликује их задржавање старих информација и након пристизања нових. Основни облици засновани су на Хебовом учењу, које дефинише ажурирање тежина неурона. Напреднију верзију – **самоорганизујуће мапе** (*SOM*) – развио је Кохонен. Код њих се улазни подаци сликају у излазну мапу, која се ажурира тако што се измене тежине погођеног/победничког неурона, као и његових суседа.

Скриптом „som.py“ формирано је 60 различитих самоорганизујућих модела, за сваку могућу комбинацију параметара – датотеку, број кластера (заправо димензије дводимензионе мреже, с тим што се прва координата фиксира на јединицу) и меру растојања. Модели су названи по шаблону типа „<datoteka>_<k>_<mera>.joblib“. Сваки је визуелизован као пратећи PNG фајл. У датотеци „somsilhouette.txt“ сачувани су и сенка коефицијенти модела, као и мапе расподеле кластера. Модели су тренирани у десет пута више итерација (епоха) него што има инстанци у скупу, док су параметри почетне стопе учења и распона суседства (сигма) фиксирани на једну половину.

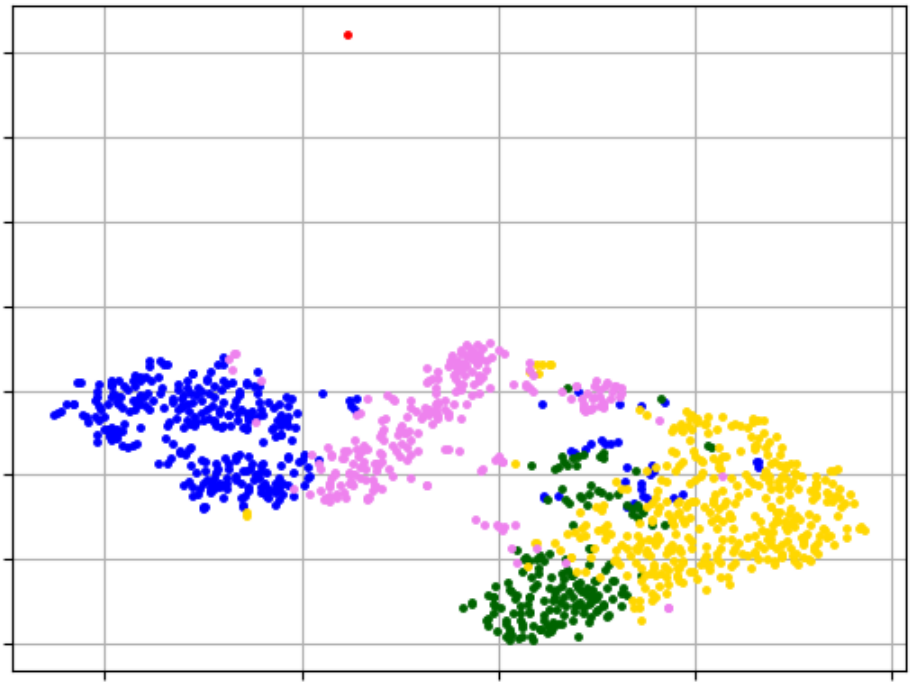
И овде су се као најбољи по сенци (и слични ранијим резултатима) показали модели са два кластера, са већ познатим проблемом неравномерне расподеле ћелија. Најравномерније резултате дао је модел са пет кластера и косинусним растојањем, а он је био и најсличнији претходним наизглед успешним косинусним 5-медијана моделом и сакупљајућим хијерархијским моделом са комплетном везом и истим бројем кластера и метриком. Његов приказ следи у наставку, кроз слике 20, 21, 22, 23 и 24.



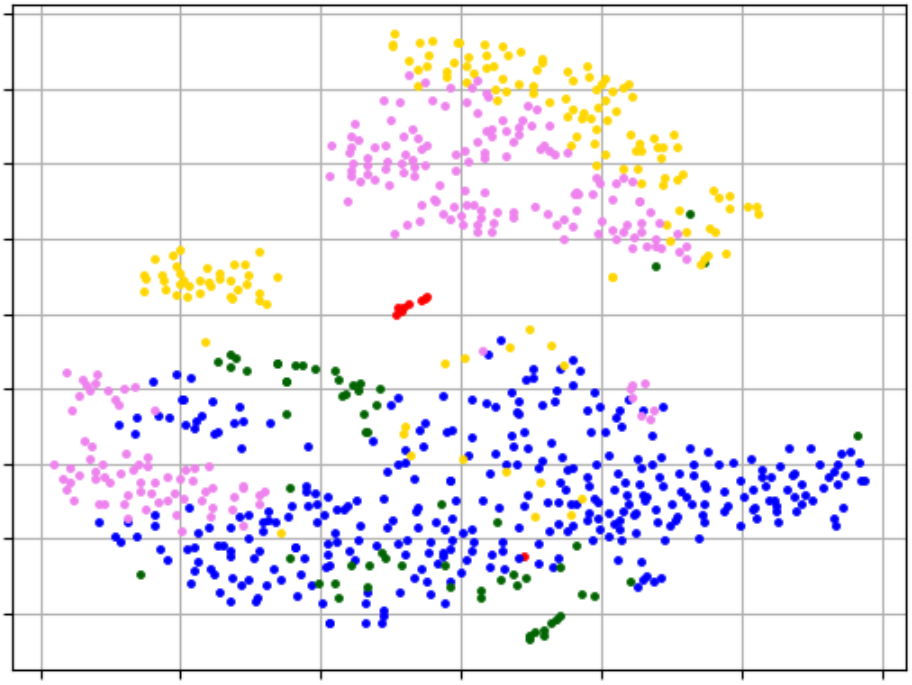
Слика 20. SOM модел са пет кластера и косинусним растојањем за прву датотеку, у сачуваним сликама назван „GSM3330561_5som_cos.png“



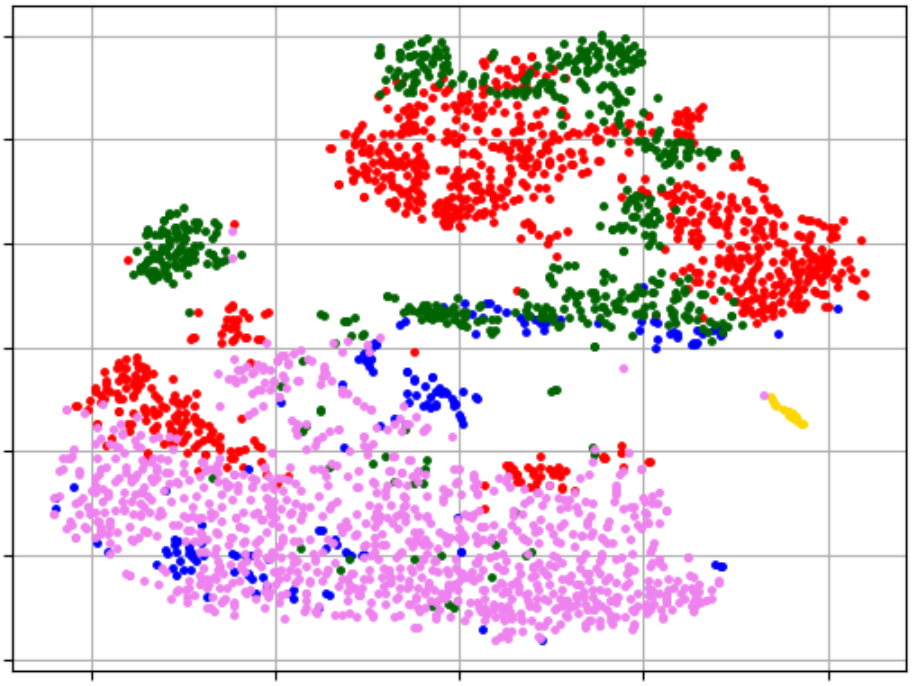
Слика 21. SOM модел са пет кластера и косинусним растојањем за другу датотеку, у сачуваним сликама назван „GSM3330562_5som_cos.png“



Слика 22. SOM модел са пет кластера и косинусним растојањем за трећу датотеку, у сачуваним сликама назван „GSM3330563_5som_cos.png“



Слика 23. SOM модел са пет кластера и косинусним растојањем за четврту датотеку, у сачуваним сликама назван „GSM3330564_5som_cos.png“



Слика 24. SOM модел са пет кластера и косинусним растојањем за спојену датотеку, у сачуваним сликама назван „GSM333056x_5som_cos.png“

Пре наставка, није лоше напоменути да, иако сличан, овај модел попут прошлог ипак одступа од косинусног 5-медијана модела. Наиме, многе ћелије нису истоветно распоређене, а и сам однос расподеле је нешто другачији. Ово се донекле може схватити као последица чињенице да су модели са репрезентативним представницима често скупљи за изградњу, па и прецизнији. То је био случај и овде: док је за прављење SOM-а или хијерархије довољно било неколико минута за фиксирани број итерација, 5-медијана модел правио се знатно дуже, са конвергенцијом као условом заустављања.

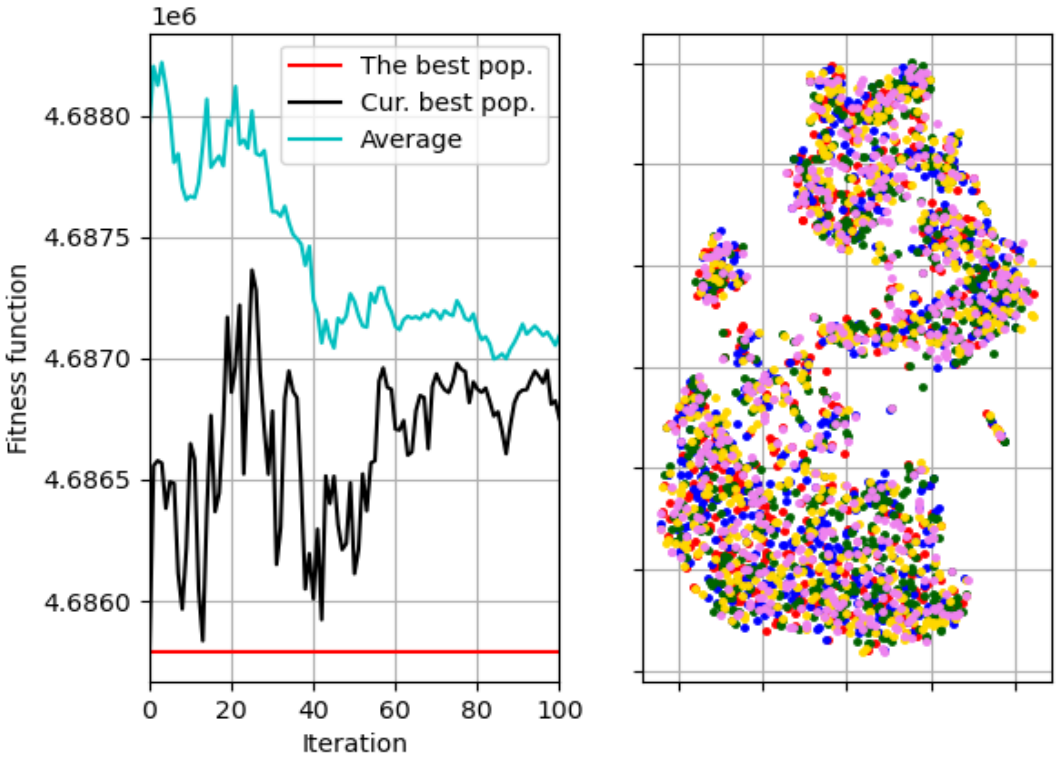
Још једна напомена могла би бити за број кластера (у конкретном случају неурона). Већ је напоменуто да су потенцијлано најкориснији модели са пет кластера, независно од сенка коефицијента, пошто би они могли бити управо РВМС. Из тог разлога се они највише разматрају, иако некад постоје незнатно бољи модели са три или четири кластера, што је могуће видети из сачуваних резултата. Ипак, све је то без осетне разлике, а без посебне могућности за тумачење, што оправдава занемаривање.

2.6. Генетски алгоритам

Следећи разматрани приступ био је **генетски алгоритам**, метахеуристички оптимизациони метод заснован на теорији еволуције, према којој углавном опстају само најбоље прилагођене јединке (хромозоми, скупови гена). Као и остала еволутивна израчунавања, генетски алгоритам припада групи метахеуристика заснованих на популацији јединки (*P*-метахеуристике), које се састоје од нпр. случајног генерисања почетне популације, а затим итеративног унапређивања исте. Решења (овде модели) из почетне популације (генерације) укрштају се, а њихови потомци, уз евентуалне мутације појединих гена, улазе у састав наредне генерације. Дакле, на почетку се случајно направи полазна генерација и евалуира се прилагођеност. Потом се, док није испуњен услов заустављања, ради следеће: одабирају се јединке из текуће популације над којима се примењују генетски оператори, укрштају се изабране јединке, потомци евентуално мутирају, нова популација се евалуира и постаје текућа. Теорема о схемама

осигурава да се, уз погодно подешене параметре, сменом генерација континуирано добија нова популација која је боље прилагођена окружењу од претходне.

Када су у питању скупови из рада, већина корака препуштена је имплементацији модула *pyclustering*. То укључује репрезентацију (кодирање) решења, функцију прилагођености (циља, погодности – мера квалитета предложеног ненадгледаног модела), генерисање почетне популације, критеријум заустављања, као и генетске операторе – селекцију, укрштање, мутацију. Од кориснички дефинисаних параметара, број кластера још једном је биран из опсега од два до пет, док је сам алгоритам радио са сто генерација од по двадесет јединки. Приступ је примењен у скрипту „genetski.py“, са моделима имена типа „<datoteka>_<k>ga.joblib“, као и пропратним визуелизацијама. Добијени резултати врло су незадовољавајући, са доста преклапања, о чему сведочи приказ модела са пет кластера из наставка („GSM333056x_5ga.png“) са слике 25. Наравно, утврђено је да та преклапања не одговарају онима из уводног разматрања, односно да на овај начин ипак није добијен модел поделе према датотекама, иако личи.



Слика 25. Генетски модел са пет кластера за спојену датотеку, у сачуваним сликама назван „GSM333056x_5ga.png“

2.7. Додатни модели

Као следеће на попису разматраних приступа налази се неколико метода кластеровања сличних већ разматраним, али са потенцијалом да дају другачије резултате. Прво међу њима јесте **просечно померање** (енгл. *mean shift*), које представља кластерирање помоћу равног кернела. Овај приступ има за циљ да у узорку открије „мрље“ равномерне густине. Алгоритам је заснован на центроидима, према чему је сличан методу *k*-средина. Важна разлика је што се кандидати на посебан начин филтрирају у фази накнадне обраде, што обично даје робуснија решења, а притом дозвољава и проглашавање појединих инстанци за елементе ван граница. На основу овога се издвајају два приступа, од којих други (онај који дозвољава аномалије)

додатно личи на метод кластеровања заснован на густини. Оба су примењена у скрипту „dodatno.py“, са моделима имена по шаблону „<datoteka>_meanshift_<pristup>.joblib“.

Друга на реду је **пропагација афинитета** (енгл. *affinity propagation*), која настоји да одреди репрезентативне представнике потенцијалних кластера из самог узорка, чиме подсећа на алгоритам *k*-медоида. Посебна је занимљивост у контексту овог рада да се пропагација афинитета показала као боља у биоинформатичким применама од класичних алгоритама из скупине оних заснованих на репрезентативним представницима. Сам алгоритам ради тако што итеративно смањује квадратну грешку, што чини еуклидску меру једином подржаном (ово, иначе, важи и за све друге методе из овог поглавља, мада би било могуће изменити имплементацију тако да прихвата још неке метрике, што, међутим, није рађено). Примењен је у скрипту „dodatno.py“, са моделима названим по уобичајеном шаблону „<datoteka>_affinity.joblib“.

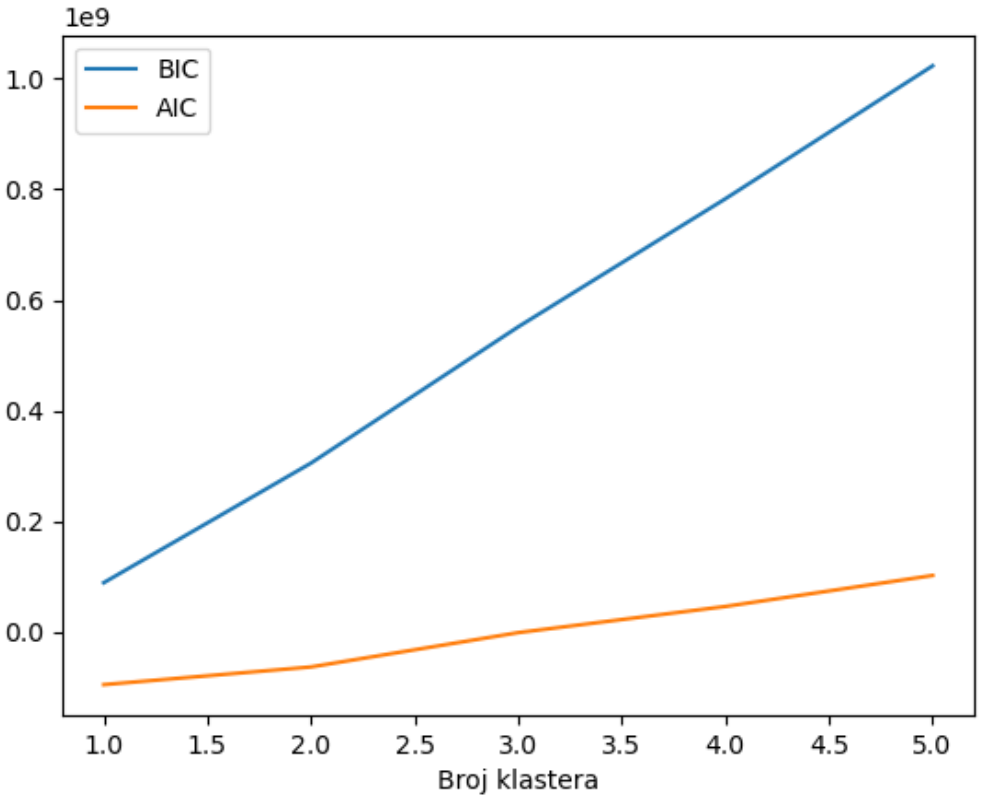
Трећи јесте алгоритам **расплинутих (фази) *c*-средина**, који по свему одговара класичном моделу *k*-средина (који је нерасплинут тј. нефази) осим по начину евалуације припадности у току извршавања алгоритма. Наиме, док код претходног приступа свака инстанца у једном тренутку припада тачно једном кластеру, у овом постоји функција припадности која одређује ком нивоу елемент припада сваком од кластера. Како је напоменуто у уводном делу, у овом раду је фокус на нерасплинутим приступима, тако да су коначни кластери тачно одређени узимањем оног са највећим степеном припадности, чиме је добијено класично нефази решење. Метод је примењен у скрипту „dodatno.py“, са моделима имена по шаблону „<datoteka>_<c>fcm.joblib“, док су сенка коефицијенти и мапа расподеле кластера складиштени у „csilhouette.txt“.

Четврти је алгоритам ***G*-средина**, који је још један од метода врло сличних *k*-средина, с тим што овде није потребно задати параметар *k*, већ је алгоритам сам у стању да погоди оптималан број кластера. Овај механизам заснован је на статистичком тесту којим се одређује да ли неки подскуп података прати Гаусову (отуд *G*) нормалну расподелу, што би указивало на то да чини један кластер. И он је примењен у скрипту „dodatno.py“, са моделима именованим по шаблону „<datoteka>_gmeans.joblib“.

По питању резултата над подацима из рада, ниједан модел се није добро показао. Код свих сем трећег, као алгоритама који не добијају податак о жељеном броју кластера, већ га сами процењују, догодило се да је сваки предложио изразито велики број група (четврти чак и по неколико стотина), што је довело до постојања великог броја малих – углавном једноцифреночланих, а доста чак и тривијалних, једночланих – кластера, као и изразито ниских сенка коефицијената. Управо због овога њихове оцене нису ни чуване. Трећи приступ довео је до релативно сличних резултата као у случају еуклидског *k*-средина кластеровања, на које личи. Како су добијени средње до веома незадовољавајући резултати, а свакако није добијена никаква нова информација, ниједан од побројаних модела није илустрован у наставку. Ипак, визуелизације резултата трећег алгоритма јесу сачуване у PNG формату, на исти начин као и досад.

За крај што се тиче додатних приступа разматран је модел **Гаусове смеше** (енгл. *Gaussian mixture*). У питању је вероватносни модел који ради под претпоставком да су подаци који се кластерују из мешавине коначног броја Гаусових расподела, односно да свака инстанца припада некој нормалној расподели са непознатим параметрима. Може се сматрати уопштењем модела *k*-средина таквим да је у причу укључена структура коваријансе података, док су центроиди заправо очекивања скривених расподела. За одређивање непознатих параметара, примењује се метод **максималне веродостојности** тј. алгоритам максимизације очекивања (EMA). Главна предност оваквог приступа огледа се у томе што отклања нагињање ка глобуларним кластерима, карактеристично за *k*-средина, те је у стању да пронађе и елипсоидне, који су одлика високе дисперзије.

За потребу анализе података о ћелијама, у скрипту „gaus.py“ направљен је по модел за сваку комбинацију датотеке и броја кластера, назван шаблоном именовања „<datoteka>_<k>gauss.joblib“. Заправо, услед количине меморије коју заузима матрица коваријанси, суштински део модела, по први пут није сачуван цео модел, већ само ознаке кластера и мере које ће ускоро бити размотрене. Похрањене су и пропратне визуелизације, које су као и досад направљене над t-SNE дводимензионом редукцијом, као и датотека „gsilhouette.txt“ са сенка коефицијентима и мапама расподеле кластера. Овога пута су додатно сачуване и визуелизоване две мере квалитета модела – **Бајесов** (BIC) и **Акаикеов информациони критеријум** (AIC). Наиме, у питању су величине које квантификују грешку модела, а главна идеја је минимизовати их. Модел са мањом вредношћу ових мера боље уклапа податке у претпостављену Гаусову расподелу, па је претпоставка да је зато бољи. У случају разматраних ћелија, вредности монотоно расту повећавањем броја кластера (визуелизација са слике 26 је за спојену датотеку), што говори да је, уопште под претпоставком нормалности, вероватније да су сви подаци из једне исте нормалне расподеле него из више различитих. Како модел који све инстанце ставља у један кластер махом није прихватљив, чини се да Гаусове смеше нису добро решење. Сами модели нису визуелизовани, пошто се ниједан посебно не издваја.



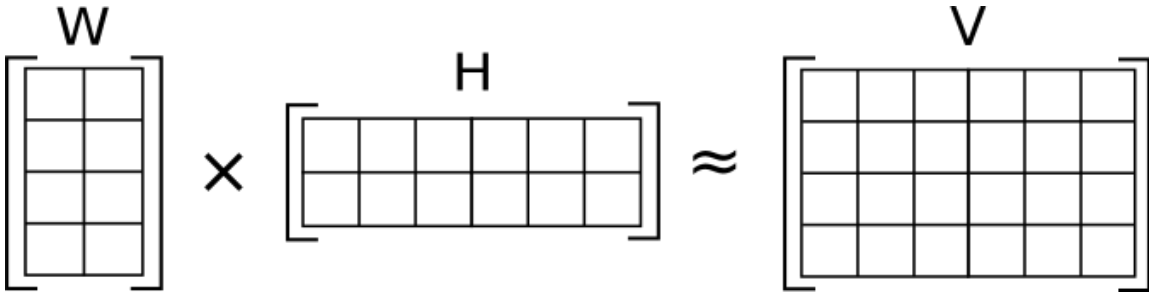
Слика 26. Вредности информационих критеријума као мере квалитета пробабилистичких модела; раст са бројем кластера указује на лоше моделе

2.8. Спектрална анализа

Последњи разматрани начин решавања задатог проблема била је **спектрална анализа**, која је раније дала добре резултате при раду са ретким подацима, углавном из области астрономије и истраживања тескта, али није занемарљив ни учинак на пољу биоинформатике. Овај тип анализе пружа други поглед на податке тако што их пре примене неког конкретног алгоритма кластеровања трансформише у нови скуп који на неки начин осликава односе унутар старог. Изворну матрицу замењује матрица сличности, а овако измењени скупови називају се **спектралним подацима**. Поменути

конкретан алгоритам углавном је *k*-средина, што и овај тип приступа чини уопштењем оних са репрезентативним представницима, па се чак понекад и назива *k*-средина са кернелом односно **кернел *k*-средина**. Ипак, услед посебне природе, као и већег значаја при раду са ретким подацима попут овде разматраних, издвојен је у ново поглавље.

Позната техника из овог домена јесте **факторизација ненегативне матрице** (енгл. *non-negative matrix factorization*, **NMF**). Улазна ненегативна матрица података V, таква да су јој редови атрибути а колоне инстанце (дакле, транспонат припремљених података из рада), разлаже се на производ две такође ненегативне матрице W и H, као у примеру са слике 27, у коме су од матрице димензија 4×6 настале две 4×2 и 2×6.



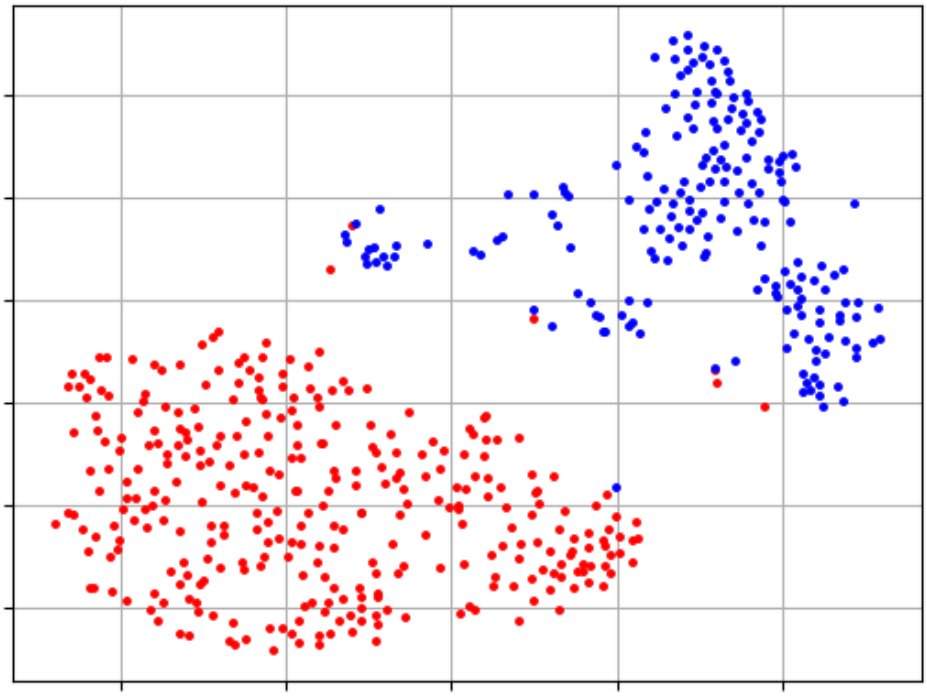
Слика 27. Пример факторизације ненегативне матрице

Добијена матрица H представља трансформисану матрицу података. Мада се главна примена овог алгоритма налази у димензионалној редукцији (средња димензија може се произвољно одабрати, а овде је димензионалност редукована са четири на два) и извлачењу атрибута (нова два атрибута представљају нове информације), по чему овај метод подсећа на анализу главних компоненти (PCA) или анализу фактора (FA), важна ствар је што је ово уједно и алгоритам кластеровања. Наиме, свака инстанца из матрице W може се сматрати кластером односно његовим центроидом (у примеру су то две тачке димензије четири), док атрибути нове матрице података H представљају степен припадности сваком од кластера. Инстанца се, очекивано, додељује кластеру са највећим степеном припадности, слично као код вероватносних или фази модела, док је сам број кластера одређен одабраном средњом димензијом тј. једнак јој је.

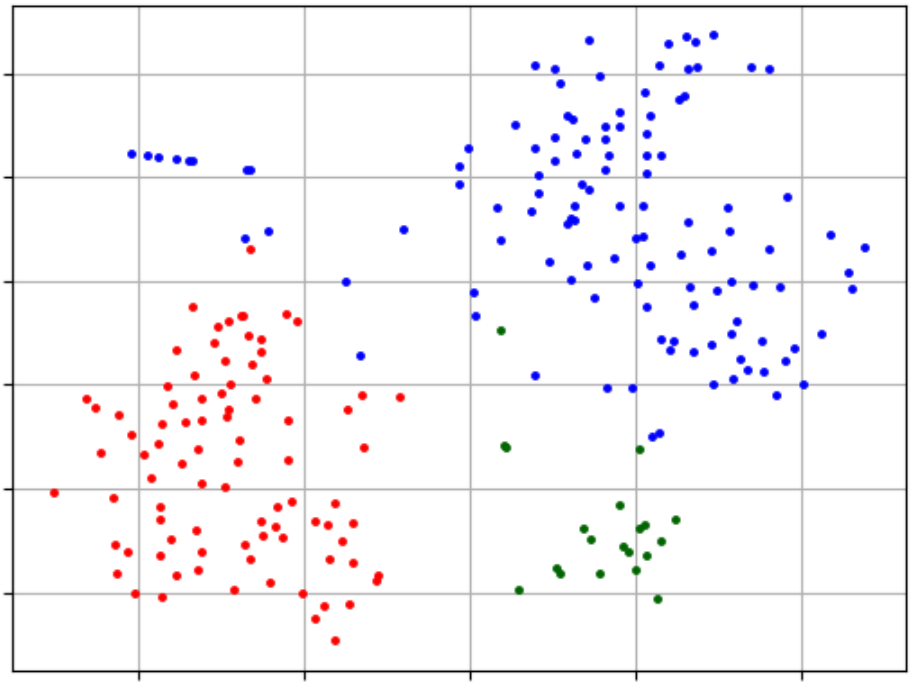
Модул *sklearn* реализује спектрално кластеровање на други начин. Улазне податке пресликава у граф повезаности тј. суседства, на основу кога даље формира Лапласову тј. Кирхофову матрицу сличности, из које је могуће дедуковати неколико важних особина графа који та матрица представља. Ово омогућава успешан рад са неконвексним подацима или, општије, у случајевима када мера растојања од центра или „распон“ кластера нису добри показатељи припадности једној групи, нпр. уколико су кластери концентрични кружни прстенови. Као начини формирања Лапласове матрице издвајају се трансформација RBF (енгл. *radial basis function*) кернелом или из графа најближих суседа. Постоје и две стратегије за додељивање ознака на основу уграђеног (трансформисаног, *embedded*) простора – *k*-средина и дискретизација. Све могуће комбинације испробане су у наставку, укључујући и мешавину *sklearn* са NMF.

Скриптом „spektralno.py“ направљено је укупно 180 спектралних модела. Модели су названи по шаблону именовања типа „<datoteka>_<k><algoritam>.joblib“. Графички приказ сваког сачуван је у PNG формату. У датотеци „ssilhouette.txt“ сачувани су и сенка коефицијенти модела, као и мапа расподеле кластера за сваки, у којој се ознака кластера слика у његову величину. Највеће сенка коефицијенте и даље имају модели са два кластера, од којих је један махом велики а други мали, те они овде нису детаљније разматрани. Добијени резултати се међусобно разликују, а они који се

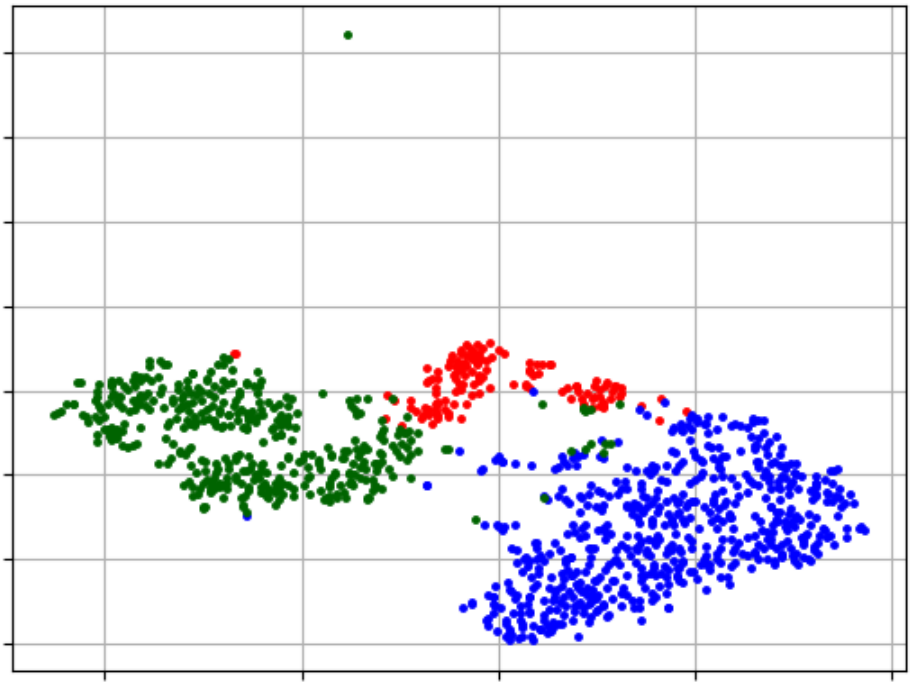
чине најбољима по визуелно чистој подели на групе издвојени су у наставку, као слике 28, 29, 30, 31 и 32. Сви илустровани модели формирали су матрицу сличности на основу графа најближих суседа, док им се стратегије додељивања ознака разликују.



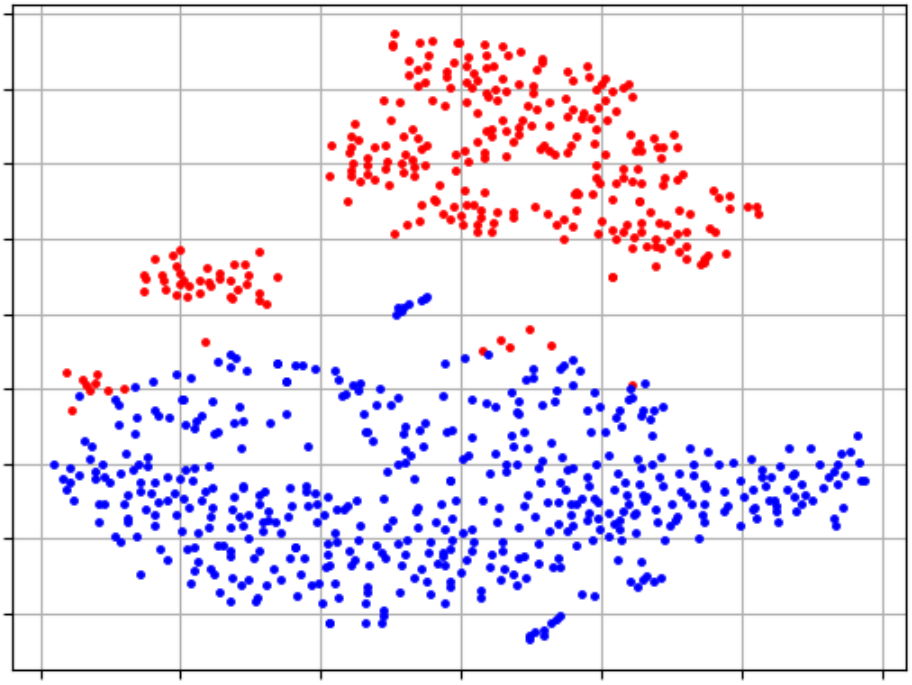
Слика 28. Спектрални модел са два кластера, најближим суседима и дискретизацијом за прву датотеку, у сачуваним сликама назван „GSM3330561_2nd.png“



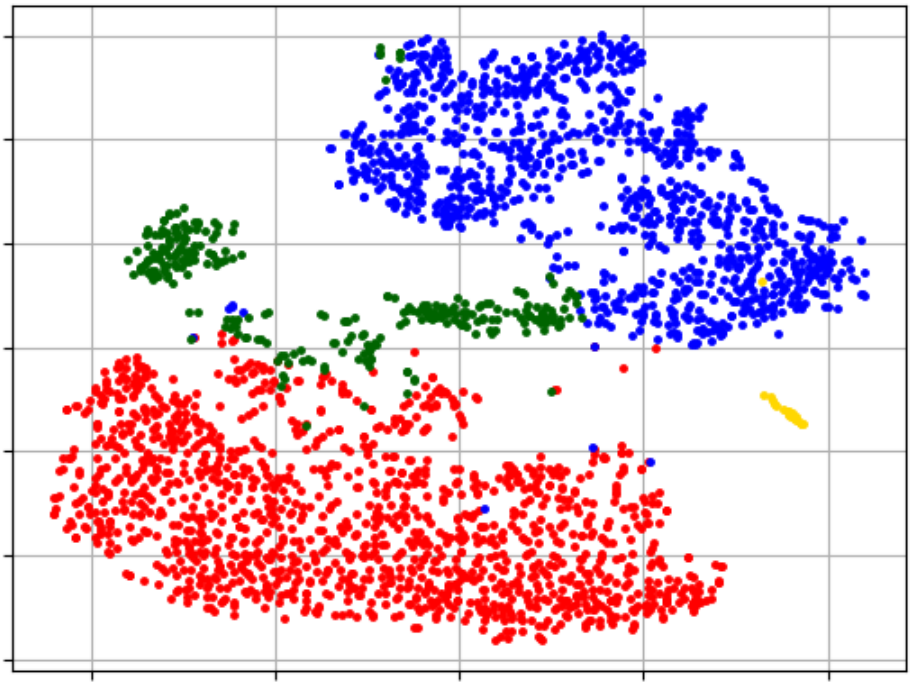
Слика 29. Спектрални модел са три кластера, најближим суседима и *k*-средина за другу датотеку, у сачуваним сликама назван „GSM3330562_3nk.png“



Слика 30. Спектрални модел са три кластера, најближим суседима и *k*-средина за трећу датотеку, у сачуваним сликама назван „GSM3330563_3nk.png“



Слика 31. Спектрални модел са два кластера, најближим суседима и дискретизацијом за четврту датотеку, у сачуваним сликама назван „GSM3330564_2nd.png“



Слика 32. Спектрални модел са четири кластера, најближим суседима и дискретизацијом за спојену датотеку, у сачуваним сликама назван „GSM333056x_4nd.png“

Када су у питању модели са пет кластера, који би потенцијално могли бити подела на пет типова PBMC ћелија, ниједан нетривијалан (у смислу да не садржи празне, једночлане или уопштено групе са врло малим бројем инстанци) нема сенка коефицијент већи од 0,2. Многима је сенка чак и негативна, што указује на веома ниску кохезију унутар кластера, као и високу блискост ћелија које нису додељене истом кластеру, што надаље указује на лош квалитет поделе. Ниједан нема расподелу ћелија по кластерима сличну претпостављеној када су у питању PBMC код људи, нити посебно личи на потенцијално добар косинусни 5-медијана модел из разматрања модела заснованих на репрезентативним представницима или њему сличне моделе добијене сакупљајућим кластеровањем или помоћу самоорганизујуће мапе. Слично је стање када се гледају модели од четири кластера за спојену датотеку, који би потенцијално могли бити подела ћелија према оригиналним датотекама пре спајања, односно према временској тачки у току терапије. У том случају сенка коефицијент није важан, пошто се на самом почетку испоставило да оваква подела није једноставна, али се и простим поређењем порекла ћелије (последња цифра GSM индекса) са добијеним групама може утврдити да формирани модели од четири кластера немају везе са тим.

2.9. Библиотека CLUTO

У додатку, посебна пажња посвећена је библиотеци за кластеровање **CLUTO** (пун назив: *CLUTO – Software for Clustering High-Dimensional Datasets*, док је сама скраћеница од *Clustering Toolkit*). Овај C-овски алат, посебно погодан за рад са великим подацима, осмислио је амерички научник Џорџ Карипис (енгл. *George Karypis*) са Одсека за рачунарство Универзитета у Минесоти још крајем прошлог века, а свој пуни потенцијал достигао је додацима из наредних година. Та проширења укључују и графичку верзију, чији кориснички интерфејс заснован на OpenGL-у олакшава рад и пружа брзе и интерактивне визуелизације, а у чијој изради су учествовали научници Мет Расмусен (енгл. *Matt Rasmussen*) и Марк Њуман (енгл. *Mark Newman*) са исте

институције. У питању је специјализована апликација за кластеровање **gCLUTO** (пун назив: *gCLUTO – Graphical Clustering Toolkit*), направљена над библиотеком, али са одређеним ограничењима; подржан је мањи број алгоритама, мера растојања, начина визуелизација и мера квалитета. Постојао је и **wCLUTO** (пун назив: *wCLUTO – Web-based Clustering of Microarray Data*), чија је идеја била да исте функционалности обезбеди корисницима са слабијим рачунарима или уопште без њих, пошто су се програми покренути на њему извршавали на серверима Центра за рачунарску геномику и биоинформатику Универзитета у Минесоти, али он већ неко време није активан.

Детаљи о самом програму и начину рада могу се наћи у документацији[8], а у наставку ће бити издвојене најважније особине и параметри које је могуће одабрати при покретању алгоритма. На првом месту је сам алгоритам, о чему ће бити речи приликом даљег рада. Следећи важан параметар је метрика, са косинусним растојањем и коефицијентом корелације (статистичка зависност инстанци) као дозвољеним вредностима. Можда најзначајнији у контексту резултата и сложености приступа јесте критеријум кластеровања, што је функција коју алгоритам настоји да оптимизује. Главни су I1, I2, E1, G1, G1′, H1, H2 (формула сваке дата је у документацији), али и уобичајена појединачна или комплетна веза за хијерархијско кластеровање. Следећи параметри омогућавају кластеровање у два корака, где се прво једним алгоритмом нађе већи број кластера од траженог, а затим се неке групе спајају хијерархијским методом док се не дође до траженог броја. У случају раздвајајућих метода, важан је и начин на који се бира кластер који се цепа. Остали параметри одређују начин рада са дрветима код хијерархијских модела, начин скалирања или нормализације података, начин прављења графа суседства код спектралних метода, број итерација, семе генератора псеудослучајних бројева итд. Постоји више параметара за контролу излазне датотеке са решењем, за визуелизације, као и за проверу квалитета добијене поделе на групе.

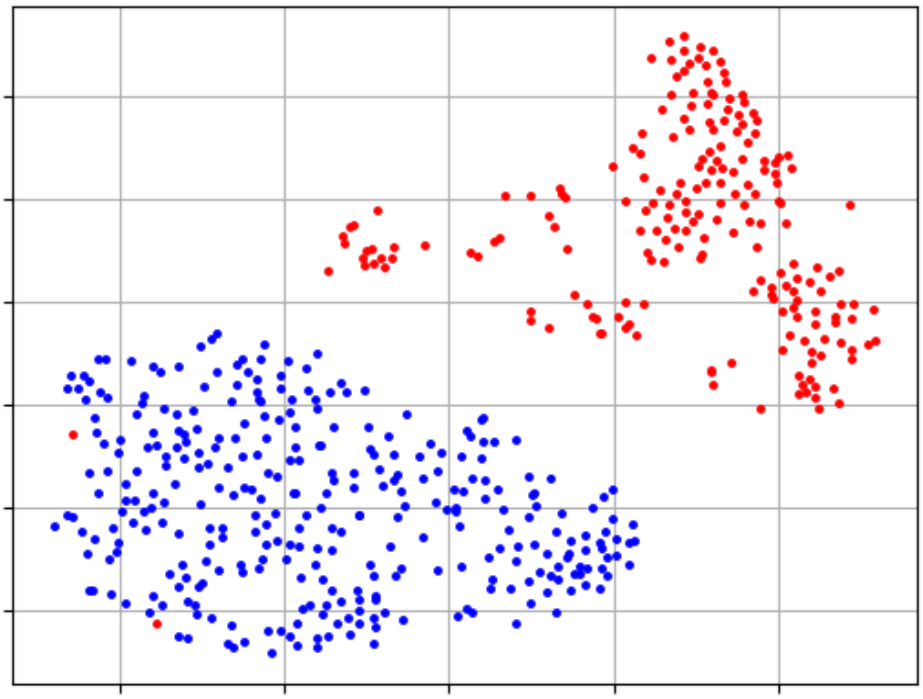
Специфичност формата улазних и излазних датотека још једна је важна одлика CLUTO-а. За разлику од досад коришћених запетом раздвојених вредности (CSV), сада се као улазни формат користи MAT, текстуални фајл чији први ред садржи димензије матрице и број ненула елемената, а сваки следећи по инстанцу исказану низом парова које чине редни бројеви ненула колона (почев од јединице) и вредности тог атрибута, раздвојени неком белином. Претварање CSV-а у MAT реализовано је скриптом „cluto.py“, при чему су резултујући фајлови названи по шаблону „<datoteka>.mat“. Основни излазни формат SOL разликује се код графичке и верзије која ради из командне линије. Код прве, у првом реду садржи три броја раздвојена табулатором – број инстанци, број кластера и грешку ако је спектрални метод. Након тога следи подела по кластерима, где први наредни ред садржи кључну реч „part“, а сваки следећи ознаку кластера инстанце коју тај ред представља. Хијерархијски модели након поделе чувају и дрволику структуру кластера, која почиње редом са кључном речју „ptree“. Код друге верзије, датотека искључиво садржи ознаке кластера, по једну у реду.

Када су у питању подаци из рада, првенствено је рађено са богатијом верзијом намењеном за рад из командне линије. Наиме, овај начин је лако било аутоматизовати у скрипту „cluto.py“, док исто не важи за графичку верзију, која захтева непрестану корисничку интеракцију. Ипак, зарад прецизније анализе и боље визуелизације најбољих модела, направљен је gCLUTO пројекат „pbmc“, у који је учитано свих пет датотека, односно њихова претходно генерисана MAT верзија. Свака је појединачно и редом кластерована, а најбољи добијени резултати у њој проучени и визуелизовани.

Први тестирани алгоритам била је **бисекција *k*-средина**, која спаја класични метод *k*-средина са раздвајајућим хијерархијским кластеровањем. Посећања ради, код хијерархијских модела постоји скуп кластера организован у облику дрвета, при чему листови садрже инстанце које се кластерују, док корен дрвета садржи целокупан улазни

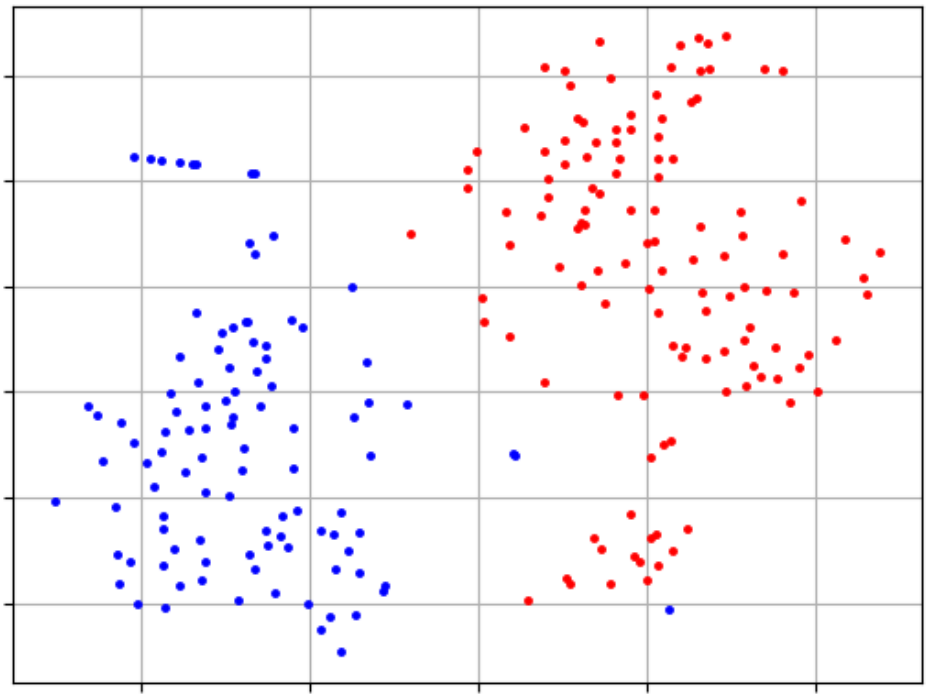
скуп. Раздвајајући приступ полази од целокупног скупа инстанци као једног кластера, дакле корена стабла, након чега итеративно дели кластере. За групу која се дели бира се нпр. највећи кластер или онај са највећим средњеквадратним одступањем тачака од центроиде. Карактеристика бисекције јесте да се у сваком кораку неки кластер дели на тачно два дела, тако да је добијено стабло бинарно. Начин поделе је метод 2-средина и он се рекурзивно извршава све док број кластера није једнак захтеваном. CLUTO имплементира како основну (под називом поновљена бисекција), тако и напреднију верзију алгоритма, која по завршетку покушава да избегне локални оптимум додатном глобалном оптимизацијом груписања. У наставку је примењен напреднији приступ.

За кластеровање скупова података о ћелијама, направљен је по модел за разне комбинације следећих параметара: број кластера од два до пет, мера косинусна и корелација, критеријум кластеровања I2, H2, E1, G1 (претходно је експериментално утврђено да се баш они најбоље понашају). Број итерација и различитих покушаја остављен је на подразумеваним вредностима, као и чињеница да се дели кластер који даје најбољу поделу. Подаци подразумевано нису нормализовани нити на неки други начин трансформисани. Од укупно сто добијених, следе модели који се издвајају. За прву датотеку, коначно је добијена визуелно добра подела на два кластера са слике 33 („GSM3330561_2rbr_cos_h2.png“). Само се за две црвене тачке чини да штрче.



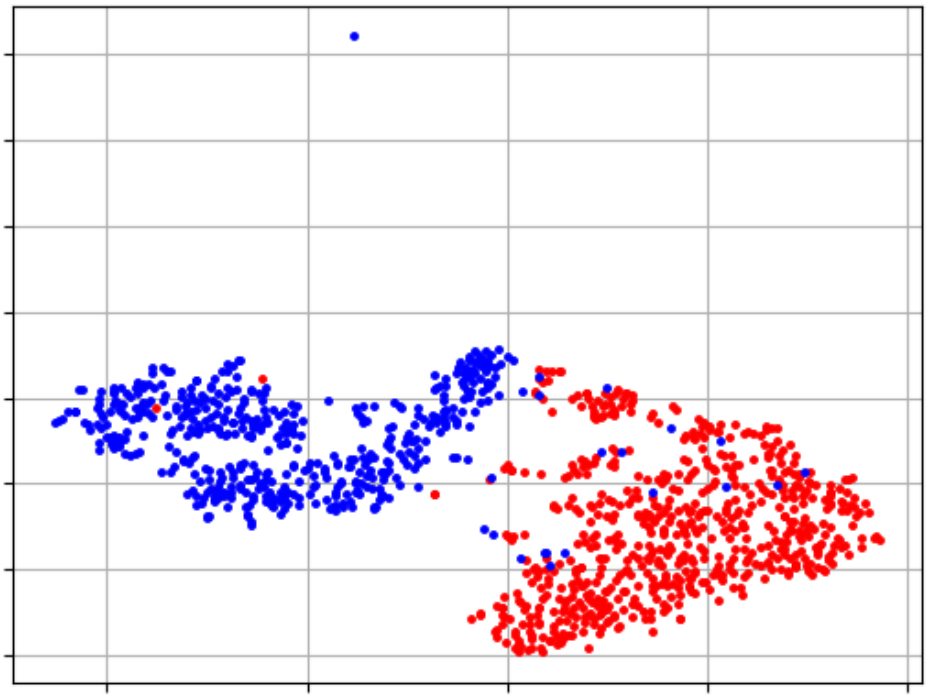
Слика 33. Поновљена бисекција са два кластера, косинусном мером и критеријумом H2 за прву датотеку, у сачуваним сликама названа „GSM3330561_2rbr_cos_h2.png“

Слично важи и за другу датотеку (слика „GSM3330562_2rbr_cos_h2.png“), где само једна плава тачка штрчи. И овај модел такође је са косинусном мером и критеријумом кластеровања H2. Илустрован је на слици 34 која следи.



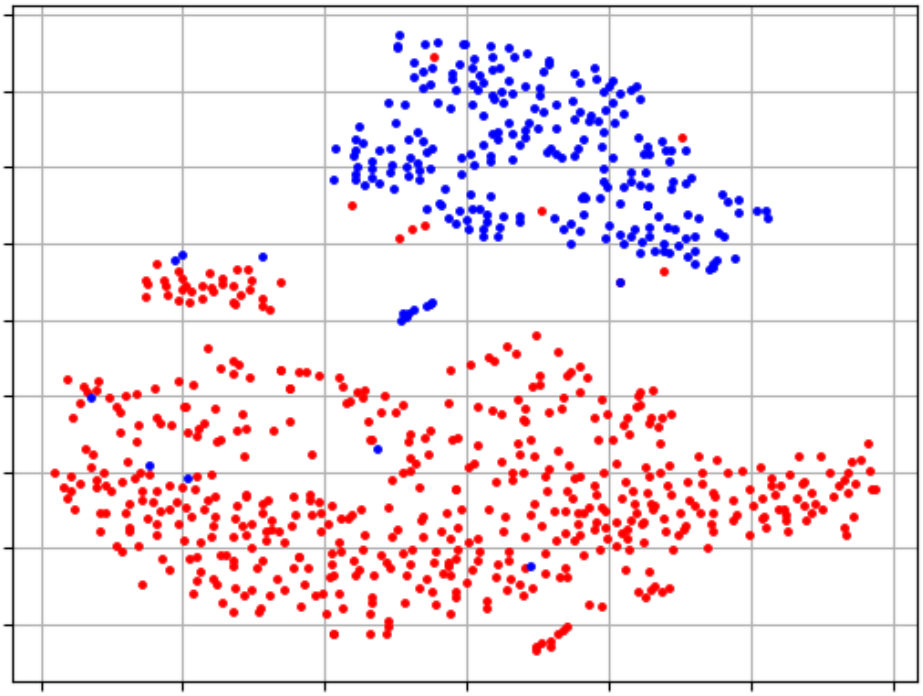
Слика 34. Поновљена бисекција са два кластера, косинусном мером и критеријумом H2 за другу датотеку, у сачуваним сликама названа „GSM3330562_2rbr_cos_h2.png“

У нешто мањој мери, ситуација се пресликала и на трећу датотеку (слика „GSM3330563_2rbr_cos_h2.png“). И овде су исти параметри, али је веће одступања. Илустрација је на слици 35 која следи у наставку.

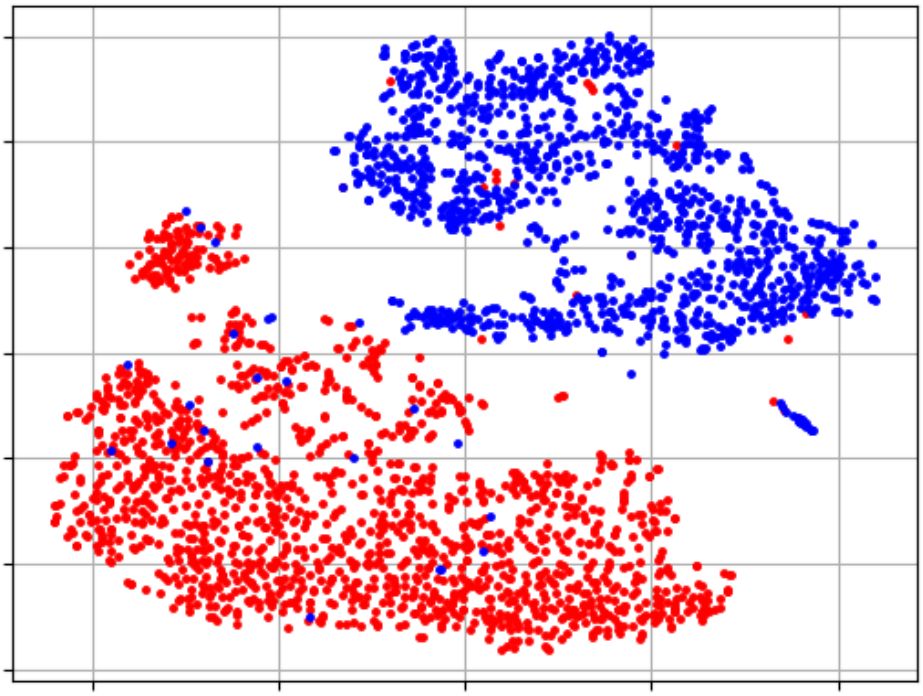


Слика 35. Поновљена бисекција са два кластера, косинусном мером и критеријумом H2 за трећу датотеку, у сачуваним сликама названа „GSM3330563_2rbr_cos_h2.png“

Све досад речено у донекле још мањој мери важи за четврту датотеку са слике 36 („GSM3330564_2rbr_cos_h2.png“), и овог пута са истим параметрима. Наиме, овде је број оступања приметно већи, што се донекле уклапа у чињеницу да је алгоритмима и досад најтеже било изборити се управо са овим скупом. Подсећања ради, једино на њему нису дали добре резултате претходно добри модели попут косинусног 5-медијана. Мало већа одступања су и код спојене са слике 37 („GSM333056x_2rbr_cos_h2.png“).



Слика 36. Поновљена бисекција са два кластера, косинусном мером и критеријумом H2 за четврту датотеку, у сачуваним сликама названа „GSM3330564_2rbr_cos_h2.png“

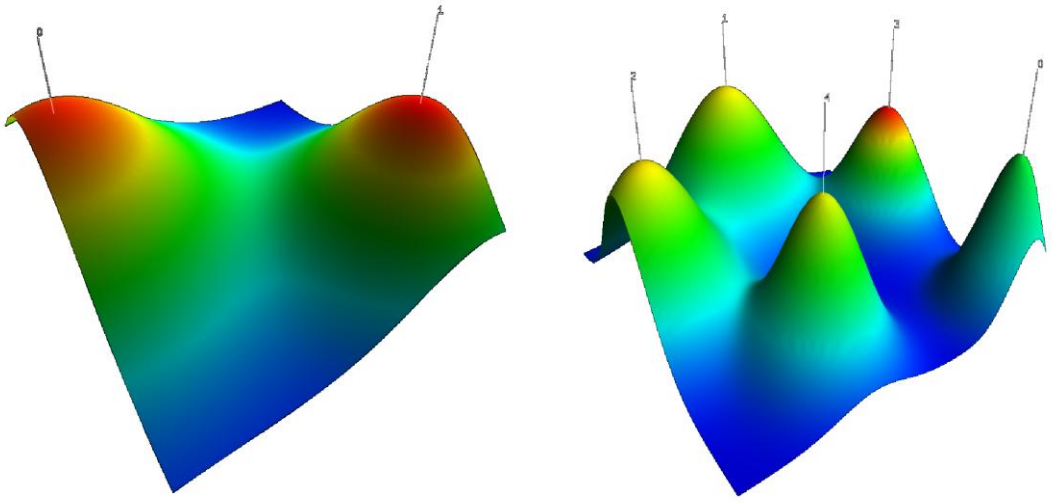


Слика 37. Поновљена бисекција са два кластера, косинусном мером и критеријумом H2 за спојену датотеку, у сачуваним сликама названа „GSM333056x_2rbr_cos_h2.png“

Све у свему, испоставља се да је бисекција дала добре резултате у првом кораку, кад је поделила инстанце у два скупа, док након тога није било посебних открића. Осим ње, CLUTO подржава и класичан алгоритам *k*-средина (под називом непосредно тј. директно кластеровање, датом на основу тога што се истовремено траже сви кластери), као и сакупљајуће хијерархијско кластеровање, али они овде нису разматрани, пошто је то већ учињено на самом почетку кластер анализе у раду. Имплементира и спектрално кластеровање (под називом графовско, датом на основу главне особине да се подаци прво преводе у граф суседства/афинитета), али оно из истог разлога није разматрано.

Ипак, последњи подржани алгоритам јесте досад неразмотрено **пристрасно сакупљајуће хијерархијско кластеровање** из два корака. Суштински је слично као основна верзија алгоритма, али се разликује по томе што пре рада проширује скуп атрибута. Наиме, прво се изврши деобно кластеровање са мањим бројем кластера, а онда се степени припадности тим кластерима, што је заправо мера сличности са центроидима, додају као нови атрибути. Тек у другом кораку гради се хијерархија. Пристрасност овог хибридног приступа огледа се у повећању сличности инстанци које су у првом кораку биле у истом кластеру, односно различитости уколико нису. Додатни критеријуми су појединачна и комплетна веза. Није било модела који се издвајају у односу на претходне, а већина заправо подсећа на резултате основног алгоритма.

За крај, поновљена је конструкција најуспешнијих модела у графичкој верзији апликације gCLUTO, за потребе приказа. За разлику од досадашње дводимензионе редукције алгоритма t-SNE, сада је примењено **вишедимензионо скалирање** (енгл. *multidimensional scaling*, **MDS**). Овиме је додата трећа димензија, у којој су кластери представљени као брда/планине (*mountain visualization*). Положај врха, његова висина, као и запремина и боја брда осликавају величину сваког кластера, њихову унутрашњу сличност и дисперзију, али и сличност са осталим кластерима. Пожељни су добро раздвојени кластери (удаљена брда), релативно сличних кардиналности (слична висина брда) и мале варијансе (боја врха што ближа црвеној). Како то изгледа, приказано је на примеру поновљене бисекције над скупом из прве датотеке, на слици 38. Лево је приказано брдо са два кластера, а десно са пет. Мада су у другом моделу брда нешто јасније просторно раздвојена, у првом се примећује осетно мања варијанса унутар кластера тј. већа унутрашња сличност, а мања спољашња, исказана црвеним врховима и мање стрмим падинама. И сенка коефицијент првог модела је бољи: 0,4 наспрам 0,3.



Слика 38. Модели поновљене бисекције прве датотеке са два и пет кластера, представљени као брда добијена вишедимензионим скалирањем

3. ЗАКЉУЧАК

Мононуклеарне ћелије периферне крви (PBMC) имају битну улогу у изучавању заразних болести и студијама у домену имунологије и аутоимунских обољења, пресађивања органа, онкологије и развоја вакцина. Користе се у проучавању рада ћелија и моделовању болести. Њиховим испитивањем се може пратити здравствено стање и дијагностификовати болести крви. Ово су само неки од разлога зашто је значајно моделовати њихово понашање и разумети механизме по којима функционишу, поготову данас када их је лако могуће секвенционирати софистицираним апаратима.

Како се кластеровање показало као најбољи начин анализе добијеног транскриптомског профила ћелија, управо њиме бави се и овај рад. Притом је главни изазов чињеница да су у питању велики и ретки подаци, који захтевају паметну и свеобухватну припрему како би добијени модели могли бити задовољавајућег квалитета. Анализирана су четири узорка особе оболеле од ретког облика карцинома, временски расподељена тако да је први узет на почетку терапије, а последњи на самом крају. Испитан је и заједнички узорак, настао спајањем сва четири улазна, са циљем провере да ли је могуће утврдити која ћелија је из ког периода лечења. Ипак, на њему нису добијени задовољавајући резултати, пошто се групе ћелија према фази терапије изразито преклапају у простору, па ниједан предложени метод заснован на уобичајеним мерама растојања није био у могућности да исправно моделује такву „неправилност“.

Пре конкретних закључака, у наставку су табеларно приказани најважнији модели и њихови сенка коефицијети. Табела је формирана скриптом „tabela.py“, који је податке прочитао из већ постојећих датотека са информацијама о сенкама, које су формиране у току прављења самих модела. Похрањена је у датотеци „tabsenke.csv“, а сумарни улазни подаци преузети су из датотеке „svesenke.txt“. Табелирани су само иоле успешни модели, док они мање успешни нису, али свакако и о њима постоје подаци у одговарајућим текстуалним датотекама. Сваки ред садржи опис модела дат бројем кластера, примењеним методом и коришћеном мером растојања (односно алгоритмом издвајања кластера у случају спектралних метода), након чега следе сенка коефицијенти добијених метода, редом за четири полазна скупа ћелија, као и за спојену датотеку означену са „*x*“. Број кластера је у распону од два до пет. Методи су именовани скраћено по називу на енглеском; нпр. *mean* означава модел *k*-средина, *comp* сакупљајући хијерархијски метод са комплетном везом, док се *nng* односи на спектрални метод са прављењем графа суседа, а нпр. *rbrh2* представља поновљену бисекцију са H2 критеријумом кластеровања. Мере имају уобичајене скраћенице, док код алгоритама спектралних метода нпр. *disc* означава дискретизацију као начин издвајања кластера из графа суседа. Ове скраћенице су већ коришћене у опису примене метода и називању генерисаних фајлова. Сенка коефицијенти су реални бројеви између -1 и 1, с тим што специјалну вредност *NaN* имају модели са једночланим кластерима. Ово је последица дељења нулом у модулу *pyclustering* приликом рачунања формуле, али није велики проблем, пошто модели са једночланим групама ионако нису пожељни.

Бр.класт.	Метод	Мера	GSM1	GSM2	GSM3	GSM4	GSMx
2	mean	man	0.47	0.4	0.58	0.44	0.43
2	mean	euc	0.21	0.86	0.49	0.23	0.88
2	mean	cos	0.4	0.66	0.28	0.29	0.68
2	medi	man	0.24	0.4	0.58	0.48	0.46
2	medi	euc	0.24	0.86	0.53	0.89	0.88
2	medi	cos	0.26	0.66	0.16	0.7	0.68
2	medo	man	0.17	0.11	0.51	0.39	0.15
2	medo	euc	0.18	0.08	0.51	0.25	0.19
2	medo	cos	0.4	0.27	0.25	0.28	0.28
3	mean	man	0.15	nan	0.55	nan	0.36
3	mean	euc	0.2	nan	0.22	0.84	0.85
3	mean	cos	0.3	0.34	0.29	0.67	0.55
3	medi	man	nan	0.35	0.59	0.49	0.39
3	medi	euc	0.19	0.37	nan	0.84	0.84
3	medi	cos	0.3	0.34	0.29	0.67	0.66
3	medo	man	0.11	0.1	0.04	0.19	0.06
3	medo	euc	0.09	0.02	0.14	0.25	0.14
3	medo	cos	0.31	0.21	0.23	0.22	0.22
4	mean	man	nan	0.1	nan	nan	0.39
4	mean	euc	nan	nan	nan	0.27	0.38
4	mean	cos	0.27	0.33	nan	0.31	0.32
4	medi	man	nan	nan	nan	nan	nan
4	medi	euc	0.14	nan	nan	0.29	0.73
4	medi	cos	0.3	0.3	nan	0.31	0.53
4	medo	man	0.07	0.1	0.1	0.01	0.05
4	medo	euc	0.13	0.07	0.12	0.12	0.13
4	medo	cos	0.31	0.2	0.25	0.21	0.17
5	mean	man	nan	nan	nan	0.37	0.28
5	mean	euc	nan	nan	nan	0.24	0.67
5	mean	cos	0.3	nan	nan	0.28	0.54
5	medi	man	nan	nan	nan	0.41	nan
5	medi	euc	nan	nan	nan	0.2	0.2
5	medi	cos	0.27	0.31	0.23	0.24	0.23
5	medo	man	-0.0	nan	0.11	0.02	0.07
5	medo	euc	0.1	nan	0.14	0.04	0.12
5	medo	cos	0.25	0.09	0.22	0.2	0.21
2	ward	euc	0.19	0.83	0.38	0.9	0.88
2	comp	man	0.7	0.4	0.66	0.46	0.59
2	comp	euc	0.72	0.86	0.84	0.92	0.92
2	comp	cos	0.24	0.63	0.7	0.7	0.68
2	aver	man	0.7	0.48	0.75	0.64	0.76
2	aver	euc	0.72	0.86	0.84	0.92	0.92
2	aver	cos	0.38	0.63	0.7	0.7	0.66
2	sing	man	0.7	0.48	0.76	0.64	0.78
2	sing	euc	0.72	0.89	0.84	0.92	0.92
2	sing	cos	0.38	0.63	0.7	0.7	0.66
3	ward	euc	0.14	0.77	0.19	0.85	0.41

3	comp	man	0.55	0.2	0.65	0.46	0.55
3	comp	euc	0.28	0.84	0.51	0.85	0.85
3	comp	cos	0.23	0.31	0.62	0.18	0.21
3	aver	man	0.58	0.47	0.74	0.55	0.74
3	aver	euc	0.57	0.84	0.72	0.84	0.85
3	aver	cos	0.38	0.37	0.62	0.36	0.62
3	sing	man	0.6	0.47	0.74	0.59	0.75
3	sing	euc	0.56	0.84	0.77	0.92	0.77
3	sing	cos	0.14	0.27	0.62	0.16	0.25
4	ward	euc	0.15	0.14	0.2	0.26	0.18
4	comp	man	0.38	0.19	0.64	0.28	0.55
4	comp	euc	0.28	0.81	0.51	0.68	0.83
4	comp	cos	0.23	0.28	0.22	0.16	0.21
4	aver	man	0.51	0.39	0.62	0.54	0.66
4	aver	euc	0.54	0.81	0.69	0.84	0.78
4	aver	cos	0.37	0.26	0.34	0.28	0.34
4	sing	man	0.58	0.39	0.71	0.51	0.73
4	sing	euc	0.5	0.81	0.72	0.84	0.71
4	sing	cos	0.01	0.18	0.27	0.15	0.25
5	ward	euc	0.1	0.14	0.2	0.26	0.18
5	comp	man	0.22	0.19	0.35	0.27	0.41
5	comp	euc	0.25	0.66	0.5	0.66	0.51
5	comp	cos	0.22	0.27	0.19	0.14	0.17
5	aver	man	0.51	0.39	0.61	0.51	0.66
5	aver	euc	0.51	0.66	0.66	0.68	0.72
5	aver	cos	0.36	0.29	0.26	0.2	0.21
5	sing	man	0.51	0.34	0.67	0.51	0.68
5	sing	euc	0.5	0.66	0.69	0.76	0.71
5	sing	cos	-0.02	0.06	0.23	0.13	0.23
2	som	man	0.49	0.36	0.58	0.44	0.49
2	som	euc	0.22	0.86	0.44	0.89	0.88
2	som	cos	0.4	0.66	0.7	0.7	0.68
3	som	man	0.23	0.34	0.51	0.42	0.45
3	som	euc	0.2	0.23	0.48	0.32	0.33
3	som	cos	0.3	0.29	0.26	0.28	0.3
4	som	man	0.22	0.21	0.42	0.34	0.42
4	som	euc	0.16	0.23	0.21	0.24	0.21
4	som	cos	0.31	0.27	0.29	0.22	0.26
5	som	man	0.14	0.28	0.42	0.31	0.41
5	som	euc	0.13	0.15	0.2	0.24	0.21
5	som	cos	0.29	0.26	0.3	0.22	0.23
2	rbf	kmean	0.02	-0.01	0.01	-0.11	-0.06
2	rbf	disc	0.0	-0.01	-0.0	-0.0	0.0
2	nng	kmean	0.18	0.08	0.14	0.15	0.14
2	nng	disc	0.19	0.09	0.14	0.15	0.14
3	rbf	kmean	-0.02	-0.21	-0.07	-0.12	-0.06
3	rbf	disc	-0.01	-0.06	-0.01	-0.05	-0.02
3	nng	kmean	0.11	0.05	0.16	0.13	0.15

3	nng	disc	0.12	0.05	0.16	0.13	0.15
4	rbf	kmean	-0.03	-0.15	-0.04	-0.17	-0.11
4	rbf	disc	-0.01	-0.14	-0.02	-0.04	-0.02
4	nng	kmean	0.09	0.03	0.12	0.06	0.16
4	nng	disc	0.11	0.06	0.13	0.11	0.16
5	rbf	kmean	-0.13	-0.21	-0.12	-0.14	-0.1
5	rbf	disc	-0.03	-0.09	-0.03	-0.09	-0.05
5	nng	kmean	0.09	0.03	0.09	0.04	0.09
5	nng	disc	0.1	0.03	0.1	0.04	0.09
2	rbrh2	cos	0.38	0.27	0.27	0.3	0.28
3	rbrh2	cos	0.35	0.18	0.21	0.21	0.22
4	rbrh2	cos	0.22	0.19	0.14	0.22	0.15
5	rbrh2	cos	0.16	0.12	0.13	0.17	0.07

Из групе модела заснованих на репрезентативним представницима, као најбољи у свим проучаваним групама (појединачне датотеке и једна спојена) показао се модел 5-медијана са косинусним растојањем. Иако су модели са два кластера давали највеће сенка коефицијенте, косинусни 5-медијана отворио је простор за потенцијално важно тумачење у светлу терапије. Према њему, највећи удео активних (могуће малигних) ћелија био је пре почетна лечења. Након тога је знатно опао, чак испод очекиваног просека за здраву особу, потенцијално због веома добре реакције на терапију, да би се годину дана касније нормализовао. Ипак, још једну годину касније активност наставља да расте, што се потенцијално може објаснити чињеницом да је у студији која прати узорке пацијент стварно стекао отпорност на терапију. Сличне резултате са истом метриком и бројем кластера постигао је сакупљајући хијерархијски модел са комплетном везом, као и самоорганизујућа мапа. Уколико је ово тумачење тачно, предложени модели – поготову први, који има највећи сенка коефицијент и најбољу расподелу – могли би се користити за проверу успешности терапије (временска серија у контексту праћења стања пацијента), па чак и откривање болести (велика активност). Управо су модели из ових група издвојени као значајни и приказани у горњој табели.

Што се тиче осталих опробаних приступа, модели засновани на густини нису довели до значајних резултата. Већина инстанци проглашавана је за аномалије или смештана у само један кластер. Ни приступи попут генетског алгоритма, просечног померања, пропагације афинитета или *G*-средина нису били добри, с тим што је код њих проблем био супротан – изразито велики број врло малих кластера. Расплинуте *c*-средине, Гаусове смеше и пристрасно сакупљање нису открили ништа ново, док је спектрална анализа била нешто боља, давши пристојне моделе са два или три кластера, па је зато и она приказана у горњој табели. Визуелно најчистију поделу на два кластера дала је бисекција *k*-средина са H2 критеријумом, па је и она табелирана. Рад са графичком апликацијом gCLUTO донео је могућност визуелизације у три димензије.

Правац даљег истраживања у смислу самог моделовања могао би бити усмерен на трансформацију (нпр. нормализацију) података пре рада. По питању провере досадашњих резултата, могло би се фокусирати на тестирање постављених хипотеза о квалитету модела, најбоље поређењем добијених ознака са резултатима класификације.

ЛИТЕРАТУРА

[1] R. A. Shaikh et al. (2019). „Classification of Five Cell Types from PBMC Samples using Single Cell Transcriptomics and Artificial Neural Networks“. 2019 IEEE International Conference on Bioinformatics and Biomedicine, стр. 2207-2213.

[2] „PBMCs – The One Stop Immune Cell Shop“ (доступно на интернет адреси https://bioscience.lonza.com/lonza_bs/CH/en/pbmcs-the-one-stop-immune-cells-shop). Immunotherapy and Hematopoietic Knowledge Center, Lonza Biosciece.

[3] GSM3330561 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3330561). National Center for Biotechnology Information, Bethesda, MD, USA.

[4] GSM3330562 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3330562). National Center for Biotechnology Information, Bethesda, MD, USA.

[5] GSM3330563 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3330563). National Center for Biotechnology Information, Bethesda, MD, USA.

[6] GSM3330564 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3330564). National Center for Biotechnology Information, Bethesda, MD, USA.

[7] Paulson, Kelly et al. (2018). „Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA“. Nature Communications. 9. 10.1038/s41467-018-06300-3.

[8] Документација за Python sklearn, MiniSom, pyclustering и (g)CLUTO.