Институт информационных и вычислительных технологий
МОН РК

Казахский Национальный Университет имени аль-Фараби

Университет Туран

Люблинский технический университет, Польша


МАТЕРИАЛЫ

III Международной научной конференции

«Информатика и прикладная математика»,

посвященная 80-летнему юбилею

профессора Бияшева Р.Г.

и 70-летию профессора Айдарханова М.Б.

26-29 сентября 2018 года, Алматы, Казахстан

Часть 2

Алматы 2018

УДК 004(063)
ББК 32.973
И74

И 74 **Информатика и прикладная математика:** Мат. III Межд. науч. конф. (26-29 сентября 2018 г). Часть 2. – Алматы, 2018. – 449 с.

ISBN 978-601-332-165-3

В сборнике опубликованы доклады, представленные по 5 секциям от Республики Казахстан, Российской Федерации, США, Латвии, Польши, Республики Беларусь, Украины, Азербайджана, Узбекистана, Японии, Кореи, Ирана, Португалии, Испании, Великобритании, Греции, Кыргызской Республики и других.

Рассмотрены актуальные вопросы в области математики, информатики и управления: математического моделирования сложных систем и бизнес-процессов, исследования и разработки защищенных и интеллектуальных информационных и телекоммуникационных технологий, математической теории управления, технологий искусственного интеллекта.

Материалы сборника предназначены для научных работников, докторантов и магистрантов, а также студентов старших курсов.

УДК 004(063)
ББК 32.973

ISBN 978-601-332-165-3

КЛАСТЕРЛІК ТАЛДАУДА ТОПТЫҚ ШЕШУДІҢ ТИІМДІ ПАРАМЕТРЛЕРІН ТАҢДАУ АЛГОРИТМІ

Черикбаева Л.Ш.[1], Калдыбекұлы Б.[2]

[1] Әл-Фарабиатындағы Қазақ Ұлттық Университеті, Алматы қаласы, Қазақстан Республикасы,
[2] Институт информационных технологий и автоматики НАН РК, г. Бишкек, Кыргызстан
e-mail: lyailya_sh@mail.ru

Аңдатпа. *Берілген жұмыста кластерлік талдауда параметрлер жиынын қабатқа бөлумен топтық шешу әдісі ұсынылды. Бастапқы берілгендердің әртүрлі параметрлері бойынша кластерлеудің L нұсқасы қалыптастырылды. Әр нұсқа үшін Дунна сапа индексі есептелінді. Кластерлеудің әр жұбы үшін Ранда индексі есептелініп, ара қашықтық матрицасы есептелінді. Нәтижесінде объектілер К кластерге бөлінді. Алынған нәтиже басқа алгоритм нәтижесімен салыстырылды.*

Кейінгі жылдарды ақпараттырды өңдеу – деректерді интеллектуалды түрде талдау бағытын зерттеу қарқынды түрде артып келеді. Деректерді талдаудың классикалық тәсілдерден айырмашылығы, бұл жерде адамның мінез-құлқын модельдеуге, жалпылама интеллектуалдық мәселелерді шешуге, үлгілерді анықтауға басты назар аударылады. Кластерлерді талдаудың негізгі мақсаты – топ ішінде бір-бірімен ұқсас болып келетін және басқа топтардан айырмашылықтары бар объектілер топтарының салыстырмалы түрде аз бөлігін бөліп алу. Талдаудың бұл түрі классификациялау есептерін шешуде ақпараттық жүйелерде, сонымен қатар деректер қорымен жұмыс жасау, интернет-құжаттарды талдау, кескіндерді сегментациялау және т.б. деректердегі заңдылықтарды анықтауда қолданылады [1-2].

Кластерлеу – бұл белгілі бір қасиеттері бар тәуелсіз бірлік ретінде қарастырылатын біртекті элементтердің белгілі бір біріктірілуін бөліп алу арқылы сегменттеу. Кластерлеу процедурасы нәтижесінде «кластерлер» пайда болады, яғни бір-біріне өте ұқсас топтар пайда болады [3].

Кластерлік талдаудың есебі (таксономия, объектілерді олардың сипаттамалары бойынша топтау, «мұғалімсіз үйрену» автоматты классификациялауы) қандай да бір (немесе жұптасқан қашықтықтағы матрица арқылы) айнымалылар жиынтығымен сипатталатын объектілер жиыны болса, кластерлер саны алдын-ала берілуі мүмкін, берілмеуі де мүмкін (соңғы жағдайда кластерлердің тиімді саны автоматты түрде анықталуы мүмкін), онда сапа өлшемі, әдетте, топ ішіндегі объектілердің шашырауына және топтар арасындағы қашықтыққа қатысты белгілі бір функционалдық ретінде қарастырылады [4].

Кластерлеу (немесе кластерді талдау) кластер деп аталатын топтарға объектілер жиынын бөлу есебі. Әрбір топта «ұқсас» объектілер болуы керек, әртүрлі топтардың объектілері мүмкіндігінше әр түрлі болуы керек. Кластерлеудің классификациялаудан ерекшелігі топтар саны нақты анықталмаған және алгоритм жұмысының процесінде анықталады [5-8].

Кластерлік талдауды қолдану жалпы түрде келесі этаптарға жіктеледі:

1. Кластерлеуге арналған объектілерді таңдау.
2. Объектілерді таңдауда қарастырылатын айнымалылар жиынтығын анықтау. Қажет болса – айнымалылардың мәндерін реттеу.
3. Объектілер арасындағы ұқсастық мәндерін есептеу.
4. Ұқсас объектілер топтарын құру.
5. Талдау нәтижелерін ұсыну.

Кластерлік талдау әдістерінің көптеген түрі бар [3,4]. Кластерлік талдаудың

топтық шешімін құру әдісінің негізгі бірнеше бағыттары бар [5,6]: консенсустық үлестірілуге, коассоциативті матрицаға негізделген, теоретикалық-графтық әдістер. Бұл жұмыста кластерлік талдаудың топтық шешімін құру әдісінің коассоциативті матрицаға негізделген әдісі қолданылады.

Кластерлік талдау есебін құрастырайық. $A=\{a_1,\ldots,a_n\}$ – объектілер жиыны болсын. $X=\{x_1,\ldots,x_n\}$ – объектілерді сипаттау, $x_i \epsilon R^d$. А жиынын К біртекті топтарға (кластерлерге) бөлу керек. Біртектілік критериі ретінде кластерлер арасындағы арақашықтық пен кластер ішіндегі объектілердің әрқалай орналасуына байланысты берілген фунционалды аламыз.

Біртектілікті түсінудің әртүрлі тәсілдерімен, қолданудың нақты аймақтарының спецификаларын ескеруге мүмкіндік беретін әртүрлі шектеулер мен бөлулердің нұсқаларын біркелкі іздеу алгоритмдерімен (метод перебора) ерекшеленетін кластерлік талдаудың көптеген әдістері бар. Дегенменде берілген критериі бойынша тиімді бөлудің нұсқаларын іздеу есебі ереже бойынша экспоненталық еңбекқорлықтан тұрады, практикада сапаны жергілікті жақсартушы ағымдағы бөлулерді әр қадам сайын түрлендіретін жуық түрдегі итерациалық алгоритмдер қолданылады. Бұл жағдайда алгоритмнің жұмысы қолданушылардың берген кейбір параметрлері көмегімен басқарылады. Байқап отырғанымыздай, кластерлердің қажетті санын анықтаудың әртүрлі әдістері бар екен. Негізінен, бұл әдістер әртүрлі топтарға бірнеше бөлу нұсқаларын қалыптастырады, ал содан соң кейбір берілген критерилер бойынша ең жақсы нұсқасын таңдап алады.

Соңғы уақыттарда кластерлік талдауда шешімдерді топтық қабылдауға негізделген тәсілдер қарқынды түрде даму үстінде. Кластерлік талдаудың әрбір алгоритмі кейбір кіріс параметрлерінен тұрады. Мысалы, кластерлер саны, шекаралық арақашықтық және т.б. Кейбір жағдайларда алгоритм жұмысының қандай параметрлері ең жақсы екендігін белгісіз болады. Осы кезде тек бір нақты параметр ғана емес, бірнеше әртүрлі параметрлері бар алгоритмдерді қолдану керек. Топтық (ансамблдік) тәсіл кластерлеу сапасын жақсартуға мүмкіндік береді. Кластерлік талдауды топтық шешуді құру әдістерінің негізгі бірнеше бағыттары бар: келісе отырып үлестіруге, коассоциативті матрицаға, үлестірулер қоспасы моделіне,

теоретикалық-графтық және басқа да әдістерге негізделінген бағыттар. Бұл жұмыста коассоциативті матрицаға негізделген топтық кластерлік талдау әдісі қолданылатын болады. Коасоциативті матрица объектілер жұбының әртүрлі бөлу нұсқаларындағы әртүрлі кластерлерде қаншалықты жиі болатындығын анықтайды.

Орташаланған коассоциативті матрица:

$$H = \sum_{l=1}^{L} w_l H_l$$

мұндағы L – әртүрлі параметрлі алгортмдерді орындау саны;

$H_l = \left(h_l(i,j)\right)$; Егер і және j объектілері 1 – алгоритмде әртүрлі кластерлерге жататын болса $h_l(i,j) = 1$, басқа жағдайда $h_l(i,j) = 0$; $w_l \geq 0$ алгоритмдер салмағы $\sum_{l=1}^{L} w_l = 1$.

Орташаланған матрица элементі ретінде объектілер арасындағы жұптық арақашықтықтар қарастырылуы мүмкін: элемент мәні жоғары болған сайын, сәйкес жұптар әртүрлі кластерлерде болуы жиірек. Қорытынды келісілген кластерлік бөлуді алу үшін кез-келген жұптық ара қашықтық матрицасына негізделген кластерлік талдаудың кез-келген алгоритмін, мысалы, иерархиялық топтауды құрудың агломеративті алгоритмін (дендограммалар құру) қолдануға болады. Топтық шешімде алгоритмнің салмағын анықтау әр-түрлі әдіспен жүреді. Берілген жұмыста топтық шешімді құру үшін алгоритм салмағын анықтау үшін кластеризациялаудың сапа индексін ескеру ұсынылады.

Кластерлік талдаудың біртекті топтық шешім алгоритмдерін анықтау үшін кейбір белгілеулер енгізейік.

A=(a1,…, an) – берілген объектілер;

X=(x1,…, xn) – объектілерді сипаттау, xi ∈ Rd

μ - кластерлік талдау алгоритмі

L –

P1,…, PL – параметрлерімен (мысалы, кластерлер санымен) μ алгоритмі арқылы алынған кластеризация.

Орташаланған коассоциативті матрица келесі формуламен беріледі:

$$H_{\text{орт}}(i,j) = \sum_{l=1}^{L} \gamma_l h_l(i,j) \qquad (1.1)$$

мұндағы, h_l - l параметрлерімен μ алгоритмі үшін алынған коассиативті матрица. γ_l − алгоритм салмағы ретінде Дунна сапа индексін алуға болады:

$$ID(P) = \frac{\frac{min}{C_k \in \text{P}} \min\limits_{C_l \in \frac{P}{C_K}} (q(C_{k,} C_{l,}))}{\max\limits_{C_k \in \text{P}} (\triangle\,(\in))}$$

$$q\left(C_{k,}C_{l,}\right) = \min_{a_i \in C_k} \min_{a_j \in C_l} (\rho(X(a_i), (X\left(a_j\right))) \quad (1.2)$$

$$\triangle (C_k) = \max_{a_i, a_j \in C_k} (\rho(X(a_i), X\left(a_j\right)))$$

Айталық, Ω параметрлер жиынында алгоритм ұқсас кластерлерге бөлетін біртекті ішкі жиындар болсын, сонымен қатар *M* ұқсас ішкі жиындар (қабаттар) болсын. Осы біртекті ішкі жиындарды табайық. Ол үшін кластеризацияның әр параметрі үшін алынған Ранда жұптық индексін есептей отырып, жұптық "жақындықты" есептейік:

$$IR\left(P_i, P_j\right) = \frac{A+D}{G}, i, j = 1,\ldots,\text{L} \quad (1.3)$$

мұндағы

A – P_i, және P_j үшін бірдей кластерлерге жататын объектілер жұбының саны;

D – әртүрлі кластерлерге кіретін объектілер жұбының саны;

G – барлық жұптар саны.

IR матрицасы ара қашықтық матрицасы болып табылады. Қажетті *M* кластерлер санымен дендограмма құру әдісін алынған жұптық ара қашықтық матрицасына қолданамыз. Іздеп отырған қабатқа бөлуді аламыз.

(1.1) коассоциативті матрица қабатқа бөлуді ескергенде қалай өзгеретінін қарастырайық. Салмақ-коэффициентінен басқа әр кластерлеуге әр қабат үшін салмақ-коэффициентін енгіземіз. Орташаланған коассоциативті матрицаның өзгертілген түрі келесідей болады:

$$H_{\text{орт}}(i,j) = \sum_{m=1}^{M} \frac{\alpha_m}{L_m} \sum_{l=1}^{L} \gamma_{l,m} h_{l,m}(i,j) \quad (1.4)$$

мұндағы, $\alpha_m \geq 0$ – m қабатының коэффициенті, $\sum_{m=1}^{M} \alpha_m = 1$;

L_m – m қабатындағы параметрлер саны.

Сонымен, ұсынылған қабатқа бөлінген кластерлік біртекті топтық шешім алгор

итмін сипатайық.

Кіріс деректері:

N – объектілер саны;

L – әртүрлі параметрлерімен k-means алгоритмінің орындалу саны;

M – параметрлер жиынындағы қабаттар саны;

X матрицасы – мөлшері n×d болатын объектілерді сипаттау;

1 – этап:

1-қадам:

k-means алгоритмін Ω – жиынынан алынған кездейсоқ параметрлермен L рет орындауға жіберу. Кластеризацияның L нұсқасын қалыптастыру.

2-қадам:

(1.2) формуласы бойынша әр кластеризация үшін Дунна сапа индексін есептеу.

3-қадам:

(1.3) формуласы бойынша кластеризацияның әрбір жұбы үшін Ранда индексін есептеп, алынған мәндермен ара қашықтық матрицасын құру керек.

4-қадам:

Кластеризацияны дендограмма құру алгоритмі көмегімен М қабатқа бөлу.

2 – этап:

(1.4)-формуласы бойынша орташаланған коассоциативті матрицасын құру. Бұл матрицаны ара қашықтық матрицасы ретінде қолданып, дендограмма құру алгоритмі көмегімен қорытынды *K* кластерлерге бөлу керек.

Нәтижесінде: Берілген объектілерді *K* кластерлерге бөлуді алу.

Эксперимент жүргізу үшін L шамасы $\sqrt{N}$ - ге жақын бүтін сан ретінде берілді. М мәні әртүрлі мәндер жиынынан алынды. Кластеризация нәтижесі Ранда индексі көмегімен бағаланды (есептелініп алынған кластеризация бастапқымен салыстырылды). Алгоритм 100 рет жүргізілді, одан соң сапа индексінің нәтижесі орташаланды. Нәтижелері көпшілікке танымал k-means алгоритмі нәтижесімен салыстырылды.

1 – кесте. Ранда индексінің нәтижесі

Алгоритмдер	Берілген деректер		
	M=L/5	M=L/4	M=L/3
k-means	0,557	0,558	0,558
топтық шешім алгоритмі	0,877	0,889	0,892

Алынған нәтижелердің көрсетуі бойынша топтық шешім алгоритмі k-means алгоритміне қарағанда дәлірек нәтиже беретінін байқауға болады. Топтық шешім алгоритмі нәтижелер дәлдігін, тұрақтылығын арттырады. Нақты шарттарды орындауда алгоритмдерді топтау сапасы артады.

Әдебиеттер

1. Б. Бериков, Г.С. Лбов. Современные тенденции в кластерном анализе. http://biocomparison.ucoz.ru/_ld/0/49_berikov_lbov.pdf.

2. Б. Бериков. Коллективалгоритмов с весами в кластерноманализе разнородных данных. // Вестник Томского Государсвенного Университета. 2013. № 2(23). С. 22-31.
3. Jain A.K., Dubes R.C. Algorithms for clustering data. Prentice Hall, NJ, 1988.
4. Jain A.K. Data clustering: 50 years beyond k-means // Pattern Recognition Letters. 2010.Vol. 31, N 8. P. 651-666.
5. GhoshJ., Acharya A. Cluster ensembles// Wiley Inter disciplinary Reviews: Data Mining and Knowledge Discovery. 2011. Vol. 1(4). P. 305-315.
6. Vega-Pons, S. Ruiz-Shulcloper, J.A Survey of Clustering Ensemble Algorithms. 2011. IJPRAI 25(3), 337–372.
7. Бериков В.Б. Методы кластерного анализа данных и сегментации изображений. 2015.-98с.

ЗАДАЧА ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ. ОБЗОР СОВРЕМЕННЫХ СЕРВИСОВ

Абдуллаева С.А., Мусабаев Р.Р.

Институт информационных и вычислительных технологий
КН МОН РК, Казахстан
e-mail: s.abdullayeva@ipic.kz

Аннотация. *Данная работа посвящена описанию такой задачи обработки естественного языка как извлечение именованных сущностей. В работе дана краткая историческая справка, теоретическая постановка. Дан обзор наиболее популярных сервисов для извлечения информации, в частности основной фокус был на исследовании такой библиотеки как библиотека DeepPavlov, работающая на нейронной сети, принцип работы которой был также описан в статье*

В современных информационных технологиях роль такой процедуры, как извлечение информации, всё больше возрастает из-за стремительного увеличения количества неструктурированной информации, в частности, в Интернете. При мониторинге новостных лент с помощью интеллектуальных агентов как раз и потребуются методы извлечения информации и преобразования её в такую форму, с которой будет удобнее работать позже.

Одной из подзадач извлечения информации является извлечение именованных сущностей. Распознавание именованных сущностей (NER) (также называемое идентификацией сущности или извлечением сущности) является подзадачей извлечения информации, которая направлена на поиск и классификацию

Содержание

СЕКЦИЯ 3
Технологии искусственного интеллекта. Интеллектуальные системы управления. Речевые технологии и компьютерная лингвистика. Распознавание образов и обработка изображений. Биоинформатика и биометрические системы. Человеко-машинное взаимодействие. Машинное обучение. Интеллектуальные робототехнические системы