ӘОЖ 519.872.1
ҒТАХР 28.23.25

https://doi.org/10.55452/1998-6688-2024-21-2-83-94

[1,3]Черикбаева Л.Ш.
PhD, ORCID ID: 0000-0001-8948-4205, e-mail: cherikbayeva.lyailya@gmail.com
[2]Мукажанов Н.К.
PhD, ORCID ID: 0000-0003-4835-5751, e-mail: n.mukazhanov@satbayev.university
[2]Алибиева Ж.М.
PhD, ORCID ID: 0000-0001-9565-5621, e-mail: alibievajibek@gmail.com
[1]Адилжанова С.А.
PhD, ORCID ID: 0000-0003-1768-064, e-mail: asaltanat81@gmail.com
[1]Тюлепбердинова Г.А.
ф.-м.ғ.к., доцент, PhD, ORCID ID: 0000-0002-4322-8983, e-mail: tyulepberdinova@gmail.com
[1]Сакыпбекова М.Ж.
PhD, ORCID ID: 0000-0002-6652-1357, e-mail: sakypbekova.meruyert@gmail.com

[1]әл-Фараби атындағы Қазақ ұлттық университеті,
050040, Алматы қ., Қазақстан
[2]Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университет,
050013, Алматы қ., Қазақстан
[3]Ақпараттық және есептеу технологиялары институты,
050040, Алматы қ., Қазақстан

РЕГУЛИЗАЦИЯ МЕН КОАССОЦИАЦИЯЛЫҚ МАТРИЦАНЫ ПАЙДАЛАНА ОТЫРЫП НАШАР БАҚЫЛАНАТЫН РЕГРЕССИЯ ЕСЕБІН ШЕШУ

Аңдатпа

Қазіргі уақытта машиналық оқыту теориясы мен әдістері (МТ) қарқынды дамып келеді және ғылым мен техниканың әртүрлі салаларында, атап айтқанда, өндірісте, білім беруде және медицинада көбірек қолданылуда. Нашар бақыланатын оқыту – әртүрлі ақпаратты талдаудың модельдері мен әдістерін әзірлеуге бағытталған машиналық оқыту зерттеулерінің бөлігі. Нашар бақыланатын оқу мәселесін қарастыру кезінде, модельдегі кейбір объектілер дұрыс емес белгіленген деп есептеледі. Бұл дәлсіздікті әртүрлі түсінуге болады. Нашар бақыланатын оқыту – бұл толық дұрыс белгіленген деректерді пайдаланудың орнына, толық, дәл емес немесе анық емес бақылау сигналдарын қолдану арқылы модель оқытылатын машиналық оқыту техникасының бір түрі. Нашар бақылау әртүрлі себептерге байланысты нақты мәселелерде жиі туындайды. Бұл деректерді белгілеу үрдісінің қымбат болуына, сенсорлардың дәлдігінің нашарлығына, мамандардың біліктілігінің жеткіліксіздігіне немесе адамның қателігіне байланысты болуы мүмкін. Мысалы, нашар бақылауды белгілеу краудсорсинг әдістерін қолдана отырып алынған жағдайларда жүзеге асырылады: әр объект үшін әртүрлі белгілер жиынтығы бар, олардың сапасы орындаушылардың шеберлігіне байланысты. Тағы бір мысал кескіндегі объектілерді анықтау есебі. Шектеу сызықтары объектілерді анықтау есептерінде кескінде анықталған объектілердің орны мен көлемін көрсетудің жалпы тәсілі. Жұмыста Вассерштейн метрикасын, әртүрлі жүйелеуді және ұқсастық матрицасы ретінде коассоциация матрицасын пайдалана отырып, көп мақсатты нашар бақыланатын регрессия есебін шешу алгоритмін ұсынамыз. Жұмыста орташа салмақты ұқсастық матрицасын есептеу алгоритмі де жетілдірілді. Соңында біз ұсынылған алгоритмді синтетикалық және нақты деректердегі бақылаудағы оқытудың және нашар бақыланатын оқытудың алгоритмдерімен салыстырамыз.

Тірек сөздер: Нашар бақыланатын оқыту, кластерлік ансамбль, көп мақсатты регрессия, төмен дәрежелі ұқсастық матрицасы, коассоциация матрицасы.

Кіріспе

Айталық, P_x үлестірімінен $X = \{x_1, \dots, x_n\}, x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$ берілсін, мұндағы n-үлгідегі объектілер саны, ал p-мүмкіндіктер кеңістігінің өлшемділігі. Өз кезегінде, $Y = \{y_1, \dots, y_n\}$, $y_i = (y_i^1, \dots, y_i^m)T \in \mathbb{R}^m$ мақсатты белгілер, мұндағы m-мақсатты белгілер кеңістігінің өлшемі.

Жартылай бақыланатын трансдуктивті оқыту мәселесінде

$X \times Y = \{(x_1, y_1), \dots, x_n, y_n)\}$ деректер жиыны қарастырылады, бірақ мақсатты белгілер $y_1, \dots, y_n, = Y_1 \subseteq Y$ қол жетімді деректердің аз ғана бөлігі үшін белгілі $x_1, \dots, x_{n_1} = X_1 \subseteq X$. Қалған объектілер $\{x_{n_1} 1, \dots, x_n\} = X_0 \subseteq X$ белгіленген.

Есептің қойылымы: $Y_0 = \{y_{n_1+1}, \dots, y_n\}$ белгілерін кейбір критерий бойынша мүмкіндігінше дәл болжау. Бақыланатын белгілердің белгісіздігін модельдеу үшін біз көп айнымалы қалыпты үлестіруді қолданамыз. Әрбір i-ші деректер нүктесі үшін, $i = 1, \dots, n_1$, мақсатты мүмкіндіктің y_i мәні анықталған жинақтаушы таралу функциясы (cdf) $F_i(y)$ бар Y_i кездейсоқ шамасының іске асырылуы деп болжаймыз $D_Y \subset R^m$:

$$Yi \sim N(\mu_i, \Sigma_i), \quad (1)$$

мұндағы $\mu_i \in \mathbb{R}^m$ орташа вектор, $\Sigma_i \in \mathbb{R}^{m \times m}$ коварианттық матрица, $i = 1, \dots, n$. Белгісіздіктің жалпы дәрежесін $T_i = |\Sigma_i|$ деп түсіндіруге болады: ол неғұрлым үлкен болса, белгінің белгісіздігі соғұрлым жоғары болады. Тиісінше, қатаң таңбаланған объектілер үшін $T_i \approx 0$ болады деп күтілуде. Сонымен есептің қойылымы объективті критерий бойынша $i = n_1$+1 үшін $F_i(y)$ анықтаушы.

Негізгі ережелер

Зерттеумен байланысты жұмыстар: Жұмыс [10] бір өлшемді мақсатты айнымалы жағдайында трансдуктивті тұжырымда әлсіз бақыланатын регрессия мәселесін шешу алгоритмін ұсынады. Ол дәлсіздікті модельдеу үшін бір айнымалы қалыпты үлестіруді пайдаланады: $Y_i \sim N(a_i, \sigma_i)$, мұндағы σ_i дәлсіздіктің көрсеткіші. Содан кейін көп реттік реттеуді пайдалана отырып, болжамды және нақты үлестірулер арасындағы қашықтықты азайту арқылы оңтайландыру мәселесін шешу ұсынылады. Мұнда ұқсастық матрицасын жуықтау үшін коассоциация матрицасы пайдаланылады және ұқсастық матрицасын алу үшін кластер ансамблі және ко-ассоциация матрицасын алу үшін кластер ансамблі және k-орталар алгоритмі қолданылған. Дегенмен, мұнда алгоритм көпөлшемді жағдайға жалпыланбайды. Көп мақсатты регрессияны шешу үшін әрбір мақсатты айнымалы үшін жеке модельді оқыту қажет. Бұл тәсілдің көмегімен мақсатты айнымалылар бір-бірінен тәуелсіз болатын мәселелерді тиімді шешуге болады.

[11] мақалада коассоциация матрицасы мен оны құру алгоритмінің толықтай талдауы берілген. Дегенмен ол k-орталар алгоритмінің негізгі нұсқасына сүйенеді, оның маңызды кемшіліктері бар, соның ішінде бір метрикалық опцияны пайдалану және кластерлердің сәйкес санын таңдау белгісіз. [12] авторлар k-орташа алгоритмі бойынша кластерлеу сапасына Евклидтіктен басқа көрсеткіштердің әсерін талдайды.

Материалдар мен әдістер

Ұсынылған әдіс: Айталық,

– $F^* = \{F_1^*, \ldots, F_{n_1}^*, \ldots, F_n^*\}$ ерікті көп айнымалы қалыпты cdf жиыны болсын, әрбір F_i^* (a_i, S_i) жұбы арқылы көрсетіледі;

– $F = F_1, \ldots, F_{n1}$ белгілі cdf жиыны болсын, әрбір F_i жұппен ұсынылған (μ_i, Σ_i).

Келесіде Σ_i және S_i екеуін де оң-анықталған матрицалар деп есептейміз. $\Sigma_i = \Sigma_i^{1/2}\Sigma_i^{1/2}$, $\mathbb{S}_i = \mathbb{S}_i^{1/2}\mathbb{S}_i^{1/2^T}$. Біз $\mathbb{S}_i^{1/2}$ элементтерін s_{jk}^i деп алып, ал $\Sigma_i^{1/2}$ элементтерін σ_{jk}^i деп белгілейміз.

Функционалдық міндет: Келесі оңтайландыру мәселесін қарастырайық:

$$\text{find } F^{**} = arg \min_{F^*} J(F, F^*)$$

мұндағы

$$J(F, F^*) = \sum W(F_i, F_i^*) + \gamma \sum_{x_i x_j \in X} W(F_i^*, F_j^*) W_{ij} \quad (2)$$

мұндағы W – 2-Вассерштейн метрикасы [60], $\gamma > 0$ – параметр, ал матрица $W = (W_{ij})$ деректер жиынының элементтері арасындағы ұқсастық өлшемдерін білдіреді. Екі көп айнымалы Гаусс үлестірімдері үшін $N(\mu_0, \Sigma_0)$ және N (μ_1, Σ_1), 2-Вассерштейн қашықтығы ол:

$$W(N(\mu_0, \Sigma_0), N(\mu_1, \Sigma_1)) = ||\mu_0 - \mu_1||_2^2 + ||\Sigma_0^{1/2} - \Sigma_1^{1/2}||_F^2.$$

$$\text{find } (\alpha^*, \mathbb{S}^*) = \arg\min J(\mu, \Sigma, a, \mathbb{S})$$

мұндағы

$$J(\mu, \Sigma, a, \mathbb{S}) = \sum_{x_i \in X_1} ||(\mu_i - a_i||_2^2 + ||\Sigma_i^{1/2} - \mathbb{S}_i^{1/2}||_F^2 + \gamma \sum_{x_i x_j \in X} W_{ij}(||a_i - a_j||_2^2 + ||\mathbb{S}_i^{1/2} - \mathbb{S}_i^{1/2}||_F^2) + \beta \sum_{i=1,\ldots,n} ||a_i||_2^2 + ||\mathbb{S}_i||_F^2).$$

Оңтайлы шешімді табу үшін a_i және $S_i^{1/2}, i = 1, \ldots, n$ элементтеріне қатысты (2) дифференциалдаймыз:

$$\frac{\partial J}{\partial a_{ij}} = 2(\mu_{ij} - a_{ij}) + 4\gamma \sum_{l=1,\ldots,n} W_{lj}(a_{lj} - a_{ij}) + 2\beta a_{ij}, i = 1, \ldots, n_1$$

$$\frac{\partial J}{\partial a_{ij}} = 4\gamma \sum_{l=1,\ldots,n} W_{lj}(a_{lj} - a_{ij}) + 2\beta a_{ij}, i = n_1, \ldots, n$$

$$\frac{\partial J}{\partial s_{jk}^i} = 2\left(s_{jk}^i - \sigma_{jk}^i\right) + 4\gamma \sum_{l=1,\dots,n} W_{lj}\left(s_{jk}^i - s_{jk}^i\right) + 2\beta s_{jk}^i, i = 1,\dots,n_1$$

$$\frac{\partial J}{\partial s_{jk}^i} = 4\gamma \sum_{l=1,\dots,n} W_{lj}\left(s_{jk}^i - s_{jk}^i\right) + 2\beta s_{jk}^i, i = n_1,\dots,n.$$

$\sum_i^{1/2}$ матрицаларының төменгі үшбұрыш екенін ескере отырып, жоғарыдағы барлық элементтерді түрлендіретін $\mathrm{vec}_2 : \mathbb{R}^{m\times m} \to \mathbb{R}^{\frac{m(m+1)}{2}}$ көмекші операциясын енгіземіз. Сондай-ақ, төменгі үшбұрышты матрицаны келесі операциясы арқылы вектордан алуға болады.

$\mathrm{vec}_2^{-1}: \mathbb{R}^{\frac{m(m+1)}{2}} \to \mathbb{R}^{m\times m}$. Ұқсас, vec_3 операциясы: $\mathbb{R}^{n\times m\times m} \to \mathbb{R}^{n\times\frac{m(m+1)}{2}}$

(сонымен қатар $\mathrm{vec}_3^{-1}: \mathbb{R}^{n\times\frac{m(m+1)}{2}} \to \mathbb{R}^{n\times m\times m}$) элементтері төменгі үшбұрышты матрицалар болып табылатын үш өлшемді тензорлар үшін анықталуы мүмкін. Белгілеу енгізсек:

$$Y_{1,0} = (\mu_1^T,\dots,\mu_{n_1}^T,0,\dots,0) \in \mathbb{R}^{n\times m}$$

$$\sum\nolimits_{1,0} = (\mathrm{vec}_2(\sum\nolimits_1^{\frac{1}{2}})^T,0,\dots,0) \in \mathbb{R}^{n\times\frac{m(m+1)}{2}}$$

$$\mathrm{B} = \mathrm{diag}(\beta+1,\dots,\beta+1,\beta,\dots,\beta) \in \mathbb{R}^{n\times n}.$$

Содан кейін оңтайландыру мәселесінің шешімін матрица түрінде беруге болады:

$$a^* = (B + 2\gamma L)^{-1}Y_{1,0}$$
$$\mathbb{S}^* = \mathrm{vec}_3^{-1}((B + 2\gamma L)^{-1}\Sigma_{1,0}) \quad (3)$$

мұндағы L – Лаплаций матрицасы, яғни, $L = D - W$, D бұл элементтері $\mathrm{D}_{ii} = \sum_i W_{ij}$-ға тең диагоналды матрица. Егер біз $W = VV^T$ болатындай $V \in \mathbb{R}^{n\times n}, q \ll n$ бар деп есептесек, онда

$$B + 2\gamma L = B + 2\gamma D - 2\gamma VV^T = G - 2\gamma VV^T.$$

мұндағы $G = B + 2\gamma D$. Вудбери сәйкестендіруін [16] пайдалану арқылы шешімдегі $O(n^3)$ амалдарын қабылдайтын кері $B + 2\gamma L$ операторын келесідей көрсетуге болады:

$$(G - 2\gamma VV^T)^{-1} = G^{-1} + 2\gamma G^{-1}V\,(\boldsymbol{I} - 2\gamma V^T G^{-1}V)^{-1}V^T G^{-1} \quad (4)$$

мұндағы G диагональды матрица, $I - 2\gamma VV^T G^{-1}\,V \in \mathbb{R}^{q\times q}$. Сондықтан кері мәнді орындау үшін $O(nq + q^3)$ қажет, бұл есептеулерді айтарлықтай азайтады, өйткені болжам бойынша $q \ll n$. Соңында біз келесі формуланы аламыз:

$$a^* = (G^{-1} + 2\gamma G^{-1}V\,(\boldsymbol{I} - 2\gamma V^T G^{-1}V)^{-1}V^T G^{-1}), Y_{1,0} \quad (5)$$

$$\mathbb{S}^* = vec_3^{-1}\quad (G^{-1} + 2\gamma G^{-1}V\,(\boldsymbol{I} - 2\gamma V^T G^{-1}V)^{-1}V^T G^{-1})\boldsymbol{\Sigma}_{1,0})$$

Орташа өлшенген ко-ассоциация матрицасы келесідей:

$$H = \sum_{l=1}^{r} \omega_l H_l, \tag{6}$$

мұндағы $H_1, \ldots, H_r$ жұбының осы бөлімнің бір кластеріне жататынын немесе жатпағанын көрсететін элементтері бар $P_1, \ldots, P_r$ бөлімдері үшін бірлескен ассоциация матрицалары, x_i, x_j ансамбль элементтерінің салмақтары, $\omega_l \geq 0, \Sigma \omega_l = 1.$

Бұл матрицаның төмен дәрежелі көрінісі бар:

$$H = VV^T,$$

мұндағы $V = [V_1 V_2 \ldots V_r]$ – блок матрицасы, $V_l = \sqrt{\omega_l}\, Z_l, Z_l \epsilon \mathbb{R}^{n \times K_l}$ l- ші бөлім үшін кластерді тағайындау матрицасы: partition:
$Z_l(i,k) = \mathbb{I}[c(x_i) = k], i = 1, \ldots, n,\ k = 1, \ldots, K_l$ және K_l, P_l, $K_l \ll n$ бөліміндегі кластерлер саны. Сонымен қатар H матрицасы үшін лаплациялық L матрицасын келесі түрде жазуға болатыны көрсетілген:

$$L = D' - H,$$
$$D' = diag(D'_{11}, \ldots, D'_{nn}),$$
$$D'_{ii} = \sum_{j=1}^{n} H(i,j) = \sum_{j=1}^{n} \sum_{l=1}^{r} \omega_l \sum_{k=1}^{K_l} Z_l(i,k) Z_l(j,k). \tag{7}$$

Енді (4) оңтайлы шешімді (6) ұқсастық матрицасының және (7) диагональды төмен дәрежелі матрицаны қолдану арқылы табуға болады.

Коассоциациялық матрица: Төмен дәрежелі ұқсастық матрицасын ұсыну үшін біз ұқсастық матрицасы ретінде орташа өлшенген коассоциация матрицасын қолданамыз. Дегенмен, орташа салмақты коассоциация матрицасын есептеудің стандартты алгоритмінің [17, 18] бірқатар кемшіліктері бар:

- Евклидтік метриканы қолданатын k-орталар алгоритмі тек сфералық кластерлерді таба алады, сондықтан деректердегі кейбір күрделі қатынастар кластерлеу нәтижесінде табылмауы мүмкін;
- Нәтижеге k-орталар алгоритмі үшін қажетті кластер санын таңдау да, ансамбльдегі әртүрлі бөлімдер саны да әсер етеді.

Бұл есептерді шешу үшін біз орташа салмақты ассоциация матрицасын есептеу алгоритмін жетілдіруді шештік. Біріншіден, біз k-орташа алгоритмінде қолданылатын қашықтық метрикасы бойынша бірлескен ассоциация матрицасын орташалауды ұсынамыз. Екіншіден, қажетсіз бөлімдердің әсерін азайту және ансамбль көлемін азайту үшін ансамбльде кластердің жарамдылық индексі тұрғысынан тек оңтайлы бөлімдерді пайдалануды ұсынамыз.

Мультиметриялық орташа өлшенген коассоциациялық матрица

$\{M_t\}_{t=1}^{d}$ нүктелер арасындағы қашықтық ретінде k- алгоритмінде нүктелер арасындағы қашықтық ретінде қолдануға болатын көрсеткіштер жиыны болсын, мысалы, p – ретті Минковский қашықтығы. Содан кейін осы жиынтықтағы әрбір метрика үшін кластерлік ансамбльдің көмегімен $\{P_l^{M_y}\}_{l=1}^{r^{M_t}}$ бөлімдерінің нұсқаларының ерікті жиынтығын алуға болады. Сол сияқты, әр бөлім үшін $H_l^{M_t}$ коассоциация матрицасын табуға болады [19]. Содан кейін біз мультиметриялық орташа өлшенген коассоциация матрицасын келесідей анықтаймыз:

$$H = \sum_{t=1}^{d} H^{M_t} = \sum_{t=1}^{d} \sum_{l=1}^{r^{M_t}} \omega_l^{M_t} H_l^{M_t},$$

мұндағы $\omega_1^{M_t}, \dots, \omega_r^{M_t}$ ансамбль элементтерінің салмақтары, $\omega_l^{M_t} \geq 0$, $\sum_{l=1}^{r^{M_t}} \omega_l^{M_t} = 1$ әрбір $M_t, t = 1, \dots, d$ үшін.

Бөлімнің салмағы $\omega_l^{M_t}$ тәуелді болатын кластерлеу сапасының индексі нүктелер арасындағы қашықтық ретінде таңдалған көрсеткішті қолдануы керек екенін ескеру қажет.

Сондықтан $\sum_{t=1}^{d} \sum_{l=1}^{r^{M_t}} \omega_l^{M_t} = 1$ емес, әрбір $M_t, t = 1, \dots, d$ үшін $\sum_{l=1}^{r^{M_t}} \omega_l^{M_t}$ деп есептейміз. Әрі қарай жақсарту ретінде коассоциация матрицаларын да өлшеуге болады.

Оңтайлы өлшенген орташа коассоциация матрицасы: Жалпы, әр бөлімдегі кластерлер саны гиперпараметр болып табылады. Мысалы, [3] ішінде екі түрлі параметрлер жиыны пайдаланылады:

Ансамбльдің өлшемі $r = 10$, i-ші бөлімдегі K_i кластерлерінің саны: $K_i = 2 + i, i = 1, \dots, r$:

Ансамбльдің өлшемі $r = 10$, i-ші бөлімдегі K_i кластерлерінің саны: $K_i = 100 + i, i = 1, \dots, r$.

Дегенмен бұл таңдау оңтайлы болмауы мүмкін. Сонымен, бірінші жағдайда кластерлердің саны аз бөлімдер үшін алгоритм салмақтары өте аз болуы мүмкін, бұл олардың орташа салмақты коассоциация матрицасына әсері шамалы болады дегенді білдіреді. Екінші жағдайда, кластерлердің үлкен саны бар бөлімдерді табудың жоғары есептеу күрделілігіне қоса, барлық алынған бөлімдер бір-біріне ұқсас және бірдей дерлік салмаққа ие болуы мүмкін. Сондай-ақ, екі жағдайда да кез келген критерий бойынша кем дегенде бір оңтайлы бөлімнің табылуына кепілдік берілмейді: мысалы, кластердің жарамдылық индексінің жергілікті оптимумына қол жеткізетін бөлім.

Біз оңтайлы бөлімдермен орташа өлшенген коассоциация матрицасын есептейтін басқа алгоритмді ұсынамыз. Осылайша алынған H∗ матрицасы оптималды орташа салмақты ко-ассоциация матрицасы деп аталады. Бұл матрица оңтайлы болып табылады, өйткені оны есептеуде кластердің жарамдылық индексіне сәйкес оңтайлы бөлімдер ғана қолданылады. Төменде оңтайлы орташа өлшенген коассоциациялық матрицаны қадамдар бойынша есептеу алгоритмі берілген:

Кіріс:

$\boldsymbol{X}$ – деректер жиыны.

$\boldsymbol{r}$ – кластерлік ансамбль өлшемі.

$\boldsymbol{k_{min}}$ – бөлімдегі кластерлердің минималды саны.

$\boldsymbol{k_{max}}$ – бөлімдегі кластерлердің максималды саны.

Шығыс:

H^* – орташа өлшенген оңтайлы коассоциация матрицасы.

Қадамдар:

k әр түрлі кластерлер саны бар k-means алгоритмін қолдана отырып X жиынынан $\{P_k\}_{k=k_{min}}^{k_{max}}$ бөлулер жиынын табыңыз.

$\{P_{k_i}\}_{i=1}^{r}$ бөлімдер жиыны үшін кластерлер индексінің мәндерінің жиынын $\{\omega_k\}_{k=k_{min}}^{k_{max}}$ есептеу.

$\{\omega_k\}_{k=k_{min}}^{k_{max}}$ жиынынан және $\{P_{k_i}\}_{i=1}^{r}$ сәйкес бөлімдер жиынынан $\{\omega_{k_i}\}_{i=1}^{r}$ мәндерінің ең үлкен r мәндерін таңдау.

$\{P_{k_i}\}_{i=1}^{r}$ бөлімдерінің жиыны үшін $\{H_{k_i}\}_{i=1}^{r}$ коассоциация матрицаларының жиынын есептеу.

$H^* = \sum_{l=1}^{r} \omega_{k_i} H_{k_i}$ оңтайлы орташа коассоциация матрицасын есептеу.

соңы.

Осылайша алынған оңтайлы орташа өлшенген коассоциация матрицасын мультиметриялық орташа өлшенген коассоциация матрицасын есептеу үшін пайдалануға болады:

$$H^* = \sum_{t=1}^{d} H^{*M_t} \quad (10)$$

C – ABR алгоритмі

Корреляциялық нашар бақыланатын регрессия (C- ABR) алгоритмінің үш негізгі нұсқасын тұжырымдаймыз:

- RNF: ұқсастық матрицасын есептеу үшін радиалды негіз функциясы пайдаланылады;
- TDKM: ұқсастық матрицасын есептеу үшін орташа салмақты төмен дәрежелі коассоциация матрицасs пайдаланылады;
- NDMKM: ұқсастық матрицасын есептеу үшін оңтайлы мультиметриялық орташа салмақты төмен дәрежелі коассоциация матрицасы (10) пайдаланылады.

Кіріс:

$\boldsymbol{X}$ – әлсіз бақыланатын деректер жиынтығы, $X_1 \subset X$- белгіленген объект, $X_2 \subset X$ -қате белгіленген объект, $X_3 \subset X$ – белгіленбеген объект.

a_i, Σ_i- әрбір $x_i \in X_1 \cup X_2$ үшін мақсатты үлестірулердің орташа векторлары мен коварианттық матрицалары.

TDKM нұсқасы: $r, \boldsymbol{\Omega}$ – кластерлік ансамбльдің өлшемі және k- кластерлеу құралдарына арналған параметрлер жиынтығы.

LROMCM нұсқасы: M, k – алгоритмге арналған көрсеткіштер жиынтығы, r – кластер ансамблінің өлшемі, k_{min} – жиындағы кластерлердің ең аз саны, k_{max} – жиындағы кластерлердің ең көп саны.

Шығыс:

a^*, S^* – болжамды орташа векторлар және $\boldsymbol{X}$ жиынындағы объектілер үшін мақсатты үлестірімдердің коварианттық матрицалары (таңбаланбаған объектіге арналған болжамдарды есепке алғанда).

RNF нұсқасының қадамдары:

(2) және (4) формулаларының көмегімен мақсатты үлестірулердің болжамды орташа векторлары мен коварианттық матрицаларын тікелей есептеу керек.

TDKM нұсқасының қадамдары:

$\boldsymbol{\Omega}$- ден кездейсоқ таңдалған параметрлер үшін кластерлік бөлімнің r нұсқаларын құрыңыз; орташа өлшенген коассоциация матрицасын есептеу.

Лаплациан графигін (6) және (7)-формуладан D' қолдану арқылы төмен дәрежелі кескінде табу.

(5) формуланы қолдану арқылы мақсатты үлестірулердің болжамды орташа векторлары мен коварианттық матрицаларын есептеу.

LROMCM нұсқасының қадамдары:

M жиынындағы метрикалармен және (8) және (9) арқылы r, k_{min}, k_{max} параметрлері бар оңтайлы мультиметриялық орташа салмақты коассоциация матрицасын есептеу.

Лаплациан графигін (6) және (7) -дегі D' көмегімен төмен дәрежелі кескінде табу.

(5) формуланы қолдану арқылы мақсатты үлестірулердің болжамды орташа векторлары мен коварианттық матрицаларын есептеу.

соңы.

Нәтижелер мен талқылау

Бұл бөлімде ұсынылған корреляциялық әлсіз бақыланатын регрессия (C-ABR) алгоритмінің үш нұсқасы салыстырылды. Біз MWD және MAE метрикакаларын ұсынған әлсіз бақыланатын оқыту алгоритмдерімен салыстыру кезінде пайдаланамыз:

$$\mathrm{MWD}(\mathrm{y}, y^*) = \frac{1}{n_{test}} \sum_{x_i \in X_{test}} \|\mu_i - a_i\|_2^2 + \left\| \sum\nolimits_i^{1/2} - \mathbb{S}_i^{1/2} \right\|_F^2$$

$$MAE(y, y^*) = \frac{1}{n_{test}} \sum_{x_i \in X_{test}} \|\mu_i - a_i\|_2.$$

Мультиметриялық орташа өлшенген коассоциация матрицасын есептеу үшін салмақтарды және к біз салмақтарды және кластерлердің оңтайлы санын анықтау үшін индекстік кластердің жарамдылығы ретінде әртүрлі $p \in \{1, 2, \infty\}$ және Silhouette Minkowski метрикалық p_p пайдаланамыз.

Монте-Карло модельдеуі

$N(\mu_k^*, \Sigma_k^*)$, $\mu_k^* = (8k + 1, 8k + 2, \ldots, 8k + d_x) \in \mathbb{R}^m, \Sigma_k^* = diag(1, \ldots, 1) \in \mathbb{R}^{d_x \times d_x}, d_x = 8$ және $k \in 1, 2, 3$

көп айнымалы қалыпты үлестірулер жиынынан 1000 объектілердің деректер жинағы жасалынды.

k-ші құрамдас бөліктен құрылған объектілер үшін мақсатты функция $Y_k = k + \varepsilon_k$ деп есептейміз, мұндағы ε_k кездейсоқ шама, d_y —өлшемді қалыпты таралу(үлестіру) функциясы $N(0, D_k D_k^T)$, D_k кездейсоқ төменгі-қалыпты үлестірімнен іріктелген элементтері бар үшбұрышты матрица және және $d_y = 4$.

Нашар бақылауды қамтамасыз ету үшін біз деректер жиынтығының 10% -ы қатаң белгіленген, деректер жиынтығының 20%-ы дұрыс белгіленген объектілерден, ал қалған 70%-ы белгіленбеген деп есептедік. Анық емес белгілеуді модельдеу үшін (1) формуласымен анықталған параметрлерді қолданамыз: $\Sigma_i = \Sigma_Y$, мұндағы Σ_Y белгіленген деректерге арналған мақсатты функцияның коварианттық матрицасы. Қатаң белгіленген объектілер үшін Σ_i матрицасын нөлдік матрица деп есептейміз.

C-ABR-TDKM алгоритмдері үшін r = 30 өлшемді кластерлік ансамбльді және i- ші бөлімдегі K_i кластерлер санын қолдандық:

$K_i = 2 + i$, $i = 1, \ldots, 30$. C-ABR -NDMKM алгоритмі үшін r = 10, $k_{min} = 2$ және $k_{max} = 30$ параметрлері қолданылды. Барлық алгоритмдер үшін β = 0.001 және γ = 0.001 жүйелеу коэффициенттері орнатылды. Алынған сапа көрсеткіштерінің орташа есеппен 100-ден астамы орындалды. Нәтижелер 1-кестеде берілген.

Кесте 1 – Монте-Карло симуляциясы бойынша салыстыру

Supervision type	C-ABR		
	RNF	TDKM	NDMKM
MWD	0.382	0.324	0.227

Тәжірибеде қолданылған деректер жиыны

CO / NOx деректер жинағы үшін [20] 2015 жылғы көмірқышқыл газы (CO) және азот оксидтері (NOx) шығарындыларын пайдаланамыз. Бұл деректер жиынтығында газ турбинасының сипаттамаларын сипаттайтын және 36733 бақылауды қамтитын 11 ерекшелік бар.

Деректердің 1%-ы қатаң белгіленген, 9%-ы дұрыс белгіленген және 90%-ы белгіленбеген деп есептеледі Деректер жиынтығы үлкен болғандықтан, дұрыс емес белгілеуді модельдеу үшін біз μ_i орташа векторларын және Σ_i ковариантық матрицаларын ең жақын 50 көршілері бойынша бағалаймыз. Қатаң белгіленген объектілер үшін дәл белгілеу орташа вектор μ_i ретінде пайдаланылады, Ал Σ_i нөлдік матрицаға тең. Синтетикалық деректер сияқты, жүйелеу коэффициенттері β = 0,001 және $\gamma = 0{,}001$ орнатылған. C-ABR-TDKM және C-ABR-TDKM алгоритмдері үшін i-ші бөлімдегі k_i кластерлерінің санымен $r = 30$ өлшемді кластерлік ансамбль қолданылады: $K_i = 10 + i,\ i = 1, .., 30.$

C-ABR-NDMKMалгоритмі$r = 10, k_min = 2$ $k_{min} = 2$ және $k_{max} = 50$ параметрлерімен оқытылған. Бақыланатын оқыту алгоритмдері (көп айнымалылы сызықтық регрессия (MLR) және XGBoost (XGB)) тек қатаң белгіленген объектілермен оқытылған. Деректер жиынындағы деректердің үлкен көлеміне байланысты RBF нұсқасында кері матрицаны табу есептеу ресурстарының, әсіресе жедел жадтың айтарлықтай көлемін қажет ететінін ескеріңіз. Нәтижелер 2 -кестеде ұсынылған.

Кесте 2 – CO/NOx деректер жинағындағы салыстыру

Supervision type	C- ABR			SR	
	RNF	TDKM	NDMKM	MLR	XGB
MWD	65.45	52.22	44.74	–	–
MAE	38.84	31.92	26.83	38.69	30.48

Осылайша, эксперименттердің нәтижелері ұсынылған әдістің дәлдігінің айтарлықтай жақсарғанын көрсетеді.

Қорытынды

Жұмыста трансдуктивті ортадағы шулы деректермен берілген көп мақсатты нашар бақыланатын регрессия мәселесі қарастырылды. Вассерштейн метрикасы мен мультирегуляризацияны пайдалана отырып, оңтайландыру мәселесін шешу алгоритмі ұсынылды. Оңтайландыру мәселесін шешуді жылдамдату үшін коассоциациялық матрицаны алу үшін кластерлік ансамбльді және алынған матрицаларды сығу үшін төмен дәрежелі көрсету әдісі қолданылды. Ұсынылған алгоритм оқыту кезінде белгісіз көп өлшемді белгілерді пайдалана алмайтын бар машиналық оқыту алгоритмдерінен артықшылығын көрсетті. Сондай-ақ оңтайлы мультиметриялық орташа салмақты коассоциациялық матрицаны енгізу арқылы орташа салмақты коасоциация матрицасын есептеуге бірнеше маңызды жақсартулар жасалынду. Жаңа тәсіл алгоритмнің сапасы мен тұрақтылығын айтарлықтай жақсарта алады, сонымен қатар әрбір нақты мәселені шешу үшін оңтайлы гиперпараметрлерді іздеуді жеңілдетеді.

ӘДЕБИЕТТЕР

1 Qin Q., Zhou X., Jiang Y. (2021) Prognosis Prediction of Stroke based on Machine Learning and Explanation Model, International Journal of Computers, Communications and Control, vol. 6, pp. 1–13.

2 Merembayev T., Amirgaliyeva S. & Kozhaly K. (2021) Using item response theory in machine learning algorithms for student response data. In 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), pp. 1–5. IEEE

3 Shahana I., Tri N., Xiao F.][Fu D.Y., Chen M.F., Zhang M., Fatahalian K., Ré C. The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning. Computer Science Math. Forum 2022.

4 Xiao Y., Yin Z., Liu B. (2020) A similarity-based two-view multiple instance learning method for classification. Knowl.-Based Syst., pp. 201–202.

5 Qian X., Lin C., Chen Z., Wang W. (2024) SAM-Induced Pseudo Fully Supervised Learning for Weakly Supervised Object Detection in Remote Sensing Images. Remote Sens., vol. 16, p. 1532.

6 Zheng S., Wu Z., Xu Y., Wei Z. (2024) Weakly Supervised Object Detection for Remote Sensing Images via Progressive Image-Level and Instance-Level Feature Refinement, Remote Sens., no. 16, p. 1203.

7 Tan C., Song W. (2024) Weakly Supervised Depth Estimation for 3D Imaging with Single Camera Fringe Projection Profilometry, Sensors, no. 24, p. 1701.

8 Wang Z., Zhang J., Bai L., Chang H., Chen Y., Zhang Y., Tao J. (2024) A Deep Learning Based Platform for Remote Sensing Images Change Detection Integrating Crowdsourcing and Active Learning, Sensors, no. 24, p. 1509.

9 Yang Z., Mahajan D., Ghadiyaram D., Nevatia R., Ramanathan V. (2019) Activity driven weakly supervised object detection, pp. 2912–2921.

10 Zhou Z.H. (2017) A brief introduction to weakly supervised learning. Natl. Sci. Rev., no. 5, pp. 44–53.

11 Bogachev V.I., Kolesnikov A. (2012) The Monge-Kantorovich problem: achievements, connections, and perspectives Russ. Math. Surv., 67, pp. 785–890.

12 Aggarwal C.C., Hinneburg A., Keim D.A. (2001) On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg.

13 Kaya H., Tьfekci P., Uzun E. (2019) Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS. Turk. J. Electr. Eng. Comput. Sci., 27, pp. 4783–4796.

14 Al-Mekhlafi Z., Senan E., Rassem T., Mohammed B., Makbol N., Alanazi A., Almurayziq T., Ghaleb F. (2022) Deep learning and machine learning for early detection of stroke and haemorrhage, Comput. Mater. Contin., 72, pp. 775–796.

15 Alhakami H., Alraddadi S., Alseady S., Baz A., Alsubait T. (2020) A Hybrid Efficient Data Analytics Framework for Stroke Prediction. Int. J. Comput. Sci. Netw. Secur, 20, pp. 240–250.

16 Higham N. (2002) Accuracy and Stability of Numerical Algorithms. SIAM.

17 Cherikbayeva L., Daiyrbayeva E., Yerimbetova A. Research of Cluster Analysis Methods for Group Solutions of the Pattern Recognition Problem, Proceedings – 6th International Conference on Computer Science and Engineering, UBMK 2021, pp. 494–497. Date Added to IEEE Xplore: 13 October 2021.

18 Merembayev T., Kurmangaliyev D., Bekbauov B., Amanbek Y. (2021) A Comparison of machine learning algorithms in predicting lithofacies: Case studies from Norway and Kazakhstan. Energies, vol. 14, no. 7.

19 Abdrazakuly N., Cherikbayeva L. (2023) Creation Of An Effective Image Processing Algorithm Based On An Ensemble Approach, Journal of Problem in Computer Science and Information Technologies, no. 3(1). https://doi.org/10.26577/1i32jpcsit2308

20 UC Irvine Machine Learning Repository: Gas Turbine CO and NOx Emission Data Set, 06 April 2021. https://archive.ics.uci.edu/ml/datasets/ Gas+Turbine+CO+ and+NOx+Emission+Data+Set

[1,3]Cherikbayeva L.Ch.
PhD, ORCID ID: 0000-0001-8948-4205, e-mail: cherikbayeva.lyailya@gmail.com
[2]Mukazhanov N.K.
PhD, ORCID ID: 0000-0003-4835-5751, e-mail: n.mukazhanov@satbayev.university
[2]Alibiyeva Zhibek
PhD, ORCID ID: 0000-0001-9565-5621, e-mail: alibievajibek@gmail.com
[1]Adilzhanova S.A.
PhD, ORCID ID: 0000-0003-1768-064, e-mail: asaltanat81@gmail.com
[1]Tyulepberdinova G.A.
ф.-м.ғ.к., доцент, PhD, ORCID ID: 0000-0002-4322-8983, e-mail: tyulepberdinova@gmail.com
[1]Sakypbekova M.Zh.
PhD, ORCID ID: 0000-0002-6652-1357, e-mail: sakypbekova.meruyert@gmail.com

[1]Al Farabi Kazakh National University, 050040, Almaty, Kazakhstan
[2]Satbayev University, 050013, Almaty, Kazakhstan
[3]Institute of Information and Computing Technologies, 050040, Almaty, Kazakhstan

SOLUTION TO THE PROBLEM WEAKLY CONTROLLED REGRESSION USING COASSOCIATION MATRIX AND REGULARIZATION

Abstract

Currently, the theory and methods of machine learning (ML) are rapidly developing and are increasingly used in various fields of science and technology, in particular in manufacturing, education and medicine. Weakly supervised learning is a subset of machine learning research that aims to develop models and methods for analyzing various types of information. When formulating a weakly supervised learning problem, it is assumed that some objects in the model are not defined correctly. This inaccuracy can be understood in different ways. Weakly supervised learning is a type of machine learning method in which a model is trained using incomplete, inaccurate, or imprecise observation signals rather than using fully validated data. Weakly supervised learning often occurs in real-world problems for various reasons. This may be due to the high cost of the data labeling process, low sensor accuracy, lack of expert experience, or human error. For example, labeling of poor control is carried out in cases obtained by crowdsourcing methods: for each object there is a set of different assessments, the quality of which depends on the skill of the performers. Another example is the problem of object detection in an image. Boundary lines are a common way to indicate the location and size of objects detected in an image in object detection tasks. The article presents an algorithm for solving a multi-objective weakly supervised regression problem using the Wasserstein metric, various regularizations and a co-association matrix as a similarity matrix. The work also improved the algorithm for calculating the weighted average co-association matrix. We compare the proposed algorithm with existing supervised learning and unsupervised learning algorithms on synthetic and real data.

Key words: Weakly supervised learning, cluster ensemble, multi-objective regression, low-rank similarity matrix, co-association matrix.

[1,3] Черикбаева Л.Ш.
PhD, ORCID ID: 0000-0001-8948-4205, e-mail: cherikbayeva.lyailya@gmail.com
[2] Мукажанов Н.К.
PhD, ORCID ID: 0000-0003-4835-5751, e-mail: n.mukazhanov@satbayev.university
[2]Алибиева Ж.М.
PhD, ORCID ID: 0000-0001-9565-5621, e-mail: alibievajibek@gmail.com
[1]Адилжанова С.А.
PhD, ORCID ID: 0000-0003-1768-064, e-mail: asaltanat81@gmail.com
[1] Тюлепбердинова Г.А.
канд. физ.-мат. наук, доцент, PhD, ORCID ID: 0000-0002-4322-8983,
e-mail: tyulepberdinova@gmail.com
[1]Сакыпбекова М.Ж.
PhD, ORCID ID: 0000-0002-6652-1357, e-mail: sakypbekova.meruyert@gmail.com

[1]КазНУ имени аль-Фараби, 050040, г. Алматы, Казахстан
[2]КазНИТУ имени К.И. Сатпаева, 050013, г. Алматы, Казахстан
[3]Институт информационных и вычислительных технологий,
050040, г. Алматы, Казахстан

РЕШЕНИЕ ЗАДАЧИ СЛАБО КОНТРОЛИРУЕМОЙ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ МАТРИЦЫ КОАССОЦИАЦИИ И РЕГУЛЯРИЗАЦИИ

Аннотация

В настоящее время теория и методы машинного обучения (МО) быстро развиваются и все шире используются в различных областях науки и техники, в частности в производстве, образовании и медицине. Слабо контролируемое обучение – это часть исследований в области машинного обучения, направленная на разработку моделей и методов анализа различных типов информации. При формулировании задачи обучения со слабо контролируемой обучением предполагается, что некоторые объекты в модели определены неправильно. Эту неточность можно понимать по-разному. Слабо контролируемое обучение – это тип метода машинного обучения, при котором модель обучается с использованием неполных, неточных или неточных сигналов наблюдения, а не с использованием полностью проверенных данных. Слабо контролируемое обучение часто возникает в реальных задачах по разным причинам. Это может быть связано с высокой стоимостью процесса маркировки данных, низкой точностью датчиков, недостатком опыта экспертов или человеческой ошибкой. Например, маркировка плохого контроля осуществляется в случаях, полученных методами краудсорсинга: для каждого объекта имеется набор различных оценок, качество которых зависит от мастерства исполнителей. Другой пример – проблема обнаружения объекта на изображении. Ограничительные линии – это распространенный способ указания местоположения и размера объектов, обнаруженных на изображении, в задачах обнаружения объектов. В статье представлен алгоритм решения многокритериальной задачи слабо контролируемой регрессии с использованием метрики Вассерштейна, различной регуляризации и матрицы коассоциации в качестве матрицы подобия. В работе также был усовершенствован алгоритм расчета средневзвешенной матрицы коассоциаций. Мы сравниваем предложенный алгоритм с существующими алгоритмами обучения с учителем и обучения без учителя на синтетических и реальных данных.

Ключевые слова: слабо контролируемое обучение, кластерный ансамбль, многоцелевая регрессия, матрица сходства низкого ранга, матрица коассоциации.