УДК 004.942, 519.177, 519.233.3

О.Л. КИРИЧЕНКО
Чернівецький національний університет імені Юрія Федьковича, Чернівці, Україна,
e-mail: *o.kyrychenko@chnu.edu.ua.*

ПРО ОДИН КЛАС ВИПАДКОВИХ МАТРИЦЬ

Анотація. Розглянуто методи оцінювання розподілу елементів стохастичної матриці з припущенням про експоненціальний розподіл елементів відповідної матриці суміжності графу. Описано два випадки, в першому з яких зроблено припущення про однорідність усіх вершин графу, а в другому — про неоднорідність розподілу вершин з відповідним обчисленням щільностей елементів. Для відповідних розподілів сформульовано тести перевірки гіпотези про належність двох вершин графу тому самому кластеру.

Ключові слова: випадкова матриця, штучний інтелект, дискретний ланцюг Маркова, bootstrap-метод, інформація за Фішером.

Для дослідження ділянок вебпростору використовують методи кластерного аналізу на графах, як одного із основних підрозділів штучного інтелекту, зокрема, методи перевірки гіпотези про входження вершин графу до одного кластера. При цьому основним математичним апаратом є випадкові матриці великої розмірності з додатковими обмеженнями на їхні елементи. Асимптотичні властивості власних значень випадкових матриць, які використовують для опису функціонування ділянок вебпростору, відрізняються від класичних властивостей тим, що елементи матриці невід'ємні та залежні по рядках або стовпчиках. У цьому випадку порушується основне припущення теорії випадкових матриць про незалежність елементів матриці [1].

Класичні та удосконалені методи кластеризації на графах описані в [2–6]. Наприклад, у [2] наведено новий алгоритм кластеризації незважених графів, який є удосконаленням стохастичної блочної моделі, а саме опуклу версію методу максимальної правдоподібності. Застосування цього методу дає кращі результати порівняно з методами, що використовують поліноміальні коефіцієнти у випадку масштабування розміру кластера. Цей алгоритм застосовний до відновлення багатьох класичних генеративних моделей для проведення кластеризації графів.

У [3] розглянуто основні методи кластеризації графів, означення та міри якості кластерів. Представлено глобальні алгоритми кластеризації для всього набору вершин графу і розглянуто задачу визначення кластера для конкретної вершини. Також зазначено області застосування алгоритмів кластеризації графів.

У запропонованій роботі розглянуто аналогічну задачу визначення кластера для конкретної вершини, розв'язання якої базується на статистичних тестах.

У [4] застосовано алгоритм графових нейронних мереж для кластеризації графів і отримано кращі результати для проведення класифікації вузлів порівняно з методами k-means (DGI), SBM, MinCut та ін.

Ще один підхід до проведення процесу кластеризації графів полягає у використанні поняття внутрішньокластерної щільності [5] на противагу міжкластерній розрідженості. В [5] було проведено експериментальне оцінювання різних підходів до кластеризації графів, а саме розглянуто алгоритми марковської кластеризації, Iterative Conductance Cutting, геометричної MST-кластеризації, які базуються на властивостях матриць великої розмірності. В [6] запропоновано більш повно і системно аналізувати дані в прикладних задачах та

проводити кластеризацію за різними показниками подібності для тих самих даних, а також досліджувати різні типи зв'язків між ними.

У [7, 8] основним математичним апаратом визначення оптимальної кількості кластерів у дослідженні ділянок вебпростору є випадкові матриці із додатковими обмеженнями на елементи. Як обмеження розглянуто квадратні стохастичні матриці, в яких сума елементів по рядках (або стовпчиках) дорівнює 1 та всі елементи є невід'ємними. Особливу увагу під час дослідження таких матриць приділено гауссівським ансамблям (Gaussian Ensemble), які є унітарними, ортогональними або симплектичними. Враховуючи результати, отримані у роботах із цього напряму, можна стверджувати, що сумісний розподіл власних значень у процесі розгляду нормування $\frac{1}{\sqrt{N}}$ (N — розмірність відповідної матриці) для зазначених ансамблів добре вивчений. Основна перевага гауссівських ансамблів ґрунтується на припущенні про нормальний розподіл елементів матриці та їхню незалежність.

Проте у випадку опису функціонування ділянок вебпростору елементи матриці суміжності графу повинні бути невід'ємними з ймовірністю 1, однак, умови незалежності для елементів матриці не існує. Тому асимптотичні властивості власних значень такої матриці будуть відрізнятися від класичних властивостей.

Як і в [7, 8], припустимо, що функціонування вебпростору описується графом з матрицею суміжності $A = A_{ij}$; $i, j = 1, \ldots, N$, де N — кількість розглянутих об'єктів (вебсторінок), а A_{ij} — кількість переходів із вебсторінки i на вебсторінку j, причому припускаємо що A_{ij} є випадковими величинами. Для спрощення всіх наступних викладок будемо вважвти, що матриця суміжності є симетричною, тобто $A_{ij} = A_{ji}$, та всі елементи, які розташовані вище головної діагоналі, є незалежними в сукупності випадковими величинами.

Розглянемо випадок, коли елементи матриці суміжності підпорядковані експоненціальному закону розподілу,

$$A_{ij} \sim i.i.d.\,\mathrm{Exp}\,(\lambda); \quad j \geq i, \;\; i = 1, \ldots, N, \tag{1}$$

тобто елементи матриці суміжності вище головної діагоналі є незалежними.

Основним об'єктом дослідження будуть стохастичні матриці із випадковими величинами, що визначаються співвідношеннями

$$P_{ij} = \frac{A_{ij}}{\sum\limits_{j=1}^{N} A_{ij}}. \tag{2}$$

Розглянемо спочатку розподіл елементів випадкової матриці P_{ij} із відповідними моментами першого та другого порядків.

Лема 1. Нехай виконується умова (1). Тоді розподіл кожного елемента матриці P має вигляд

$$f_{P_{ij}}(y) = (N-1)(1-y)^{N-2}, \;\; y \in (0,1).$$

Доведення. Розглянемо перетворення

$$y_1 = \frac{x_1}{x_1 + \ldots + x_N}, \quad y_2 = x_2, \ldots, y_N = x_N.$$

Відповідне обернене перетворення має вигляд

$$x_1 = y_1 \frac{y_2 + \ldots + y_N}{1 - y_1}, \;\; x_2 = y_2, \ldots, x_N = y_N.$$

Якобіан оберненого перетворення такий:

$$\frac{\partial x}{\partial y}=\begin{pmatrix} \dfrac{y_2+\ldots+y_N}{(1-y_1)^2} & \dfrac{y_1}{1-y_1} & \cdots & \dfrac{y_1}{1-y_1} \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

з визначником

$$\left|\det\left(\frac{\partial x}{\partial y}\right)\right|=\frac{y_2+\ldots+y_N}{(1-y_1)^2}$$

за умов $y_1\in(0,1)$ та $y_2>0,\ldots,y_N>0$. Враховуючи незалежність випадкових величин $A_{i1},\ldots,A_{iN}$, отримуємо

$$f_{P_{ij}}(y)=\lambda^N\int\limits_{R_+^{N-1}} e^{-\lambda\left(y_1\frac{y_2+\ldots+y_N}{1-y_1}+y_2+\ldots+y_N\right)}\frac{y_2+\ldots+y_N}{(1-y_1)^2}\,dy_2\ldots dy_N=$$

$$=\frac{\lambda^N(N-1)}{(1-y_1)^2}\int\limits_{R_+^{N-1}} e^{-\lambda\frac{1}{1-y_1}(y_2+\ldots+y_N)}\,y_N\,dy_2\ldots dy_N=$$

$$=(N-1)(1-y_1)^{N-2}\int\limits_0^\infty e^{-y_N}y_N\,dy_N=$$

$$=(N-1)(1-y_1)^{N-2},\ \ y_1\in(0,1).$$

Лема 1 доведена.

Враховуючи розподіл елемента матриці P із леми 1, отримуємо, що середнє значення та значення другого моменту кожного елемента матриці P має вигляд

$$EP_{ij}=\int\limits_0^1(N-1)(1-y)^{N-2}\,y\,dy=\frac{1}{N};$$

$$EP_{ij}^2=\int\limits_0^1(N-1)(1-y)^{N-2}\,y^2dy=\frac{2}{N(N+1)}\approx\frac{2}{N^2}.$$

Наведені результати узгоджуються із відповідними припущеннями в [7]. Також умови леми 1 відповідають припущенням щодо однорідності переходів усередині кластера, а саме той факт, що розподіл переходу на довільну вебсторінку із того самого кластера є сталим. Ця умова порушується у випадку розгляду системи із більше ніж одним кластером, а саме перехід між кластерами є зазвичай нижчий ніж у самих кластерах, тобто $\lambda_{\text{between}}<<\lambda_{\text{within}}$, де λ_{within} — інтенсивність переходу в кластері, λ_{between} — інтенсивність переходу між кластерами. Отже, враховуючи найбільш загальний випадок про довільну кількість кластерів, припускаємо, що

$$A_{ij}\sim i.i.d.\,\mathrm{Exp}(\lambda_j),\ j\ge i,\ \ i=1,\ldots,N. \tag{3}$$

Лема 2. Нехай виконується умова (3). Тоді розподіл елемента матриці P_{ij} має вигляд

$$f_{P_{ij}}(y) = \Lambda(1-y_1)^{N-2} \sum_{\substack{k=1,\\ k\neq j}}^{N} \left((\lambda_j y + \lambda_k - \lambda_k y)^{-2} \prod_{\substack{u=1,\\ u\neq k,j}}^{N} (\lambda_j y + \lambda_u - \lambda_u y)^{-1} \right) \quad (4)$$

для $y \in (0,1)$.

Доведення. Аналогічно до твердження леми 1, розглянемо перетворення

$$y_1 = \frac{x_1}{x_1 + x_2 + \ldots + x_N},\ y_2 = x_2, \ldots, y_N = x_N .$$

Обернене перетворення та Якобіан будуть визначатися, як у лемі 1. Враховуючи незалежність випадкових величин $A_{i1}, \ldots, A_{iN}$, отримуємо

$$f_{P_{i1}}(y) = \Lambda \int\limits_{R_+^{N-1}} e^{-\left(\lambda_1 y_1 \frac{y_2 + \ldots + y_N}{1-y_1} + \lambda_2 y_2 + \ldots + \lambda_N y_N\right)} \frac{y_2 + \ldots + y_N}{(1-y_1)^2} dy_2 \ldots dy_N =$$

$$= \frac{\Lambda}{(1-y_1)^2} \int\limits_{R_+^{N-1}} e^{-\sum_{j=2}^{N} y_j \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_j\right)} (y_2 + \ldots + y_N)\, dy_2 \ldots dy_N ,$$

де $\Lambda = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_N$.

Обчислимо окремо кожний доданок

$$\frac{\Lambda}{(1-y_1)^2} \int\limits_{R_+^{N-1}} e^{-\sum_{j=2}^{N} y_j \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_j\right)} y_k dy_2 \ldots dy_N =$$

$$= \frac{\Lambda}{(1-y_1)^2} \prod_{\substack{j=2,\\ j\neq k}}^{N} \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_j\right)^{-1} \int\limits_0^{\infty} e^{-y_k \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_k\right)} y_k dy_2 \ldots dy_N =$$

$$= \frac{\Lambda}{(1-y_1)^2} \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_k\right)^{-2} \prod_{\substack{u=2,\\ u\neq k}}^{N} \left(\frac{\lambda_1 y_1}{1-y_1} + \lambda_u\right)^{-1} =$$

$$= \Lambda(1-y_1)^{N-2} (\lambda_1 y_1 + \lambda_k - \lambda_k y_1)^{-2} \prod_{\substack{u=2,\\ u\neq k}}^{N} (\lambda_1 y_1 + \lambda_u - \lambda_u y_1)^{-1} .$$

Таким чином,

$$f_{P_{i1}}(y) = \Lambda(1-y_1)^{N-2} \sum_{k=2}^{N} \left((\lambda_1 y_1 + \lambda_k - \lambda_k y_1)^{-2} \prod_{\substack{u=2,\\ u\neq k}}^{N} (\lambda_1 y_1 + \lambda_u - \lambda_u y_1)^{-1} \right).$$

Аналогічно доводиться вигляд щільності $f_{P_{ij}}(y)$ для довільного j з урахуванням перетворення

$$y_j = \frac{x_j}{x_1 + x_2 + \ldots + x_N},\ \ y_1 = x_1, \ldots, y_{j-1} = x_{j-1},\ \ y_{j+1} = x_{j+1}, \ldots, y_N = x_N .$$

Це зауваження і доводить твердження (4) леми 2.

Як можна бачити із леми 2, щільність розподілу значно ускладниться порівняно з аналогічним однорідним випадком, розглянутим у лемі 1. Крім того, відношення між щільностями P_{ij} та P_{im} буде залежати від відношення між інтенсивностями та не буде залежати від спільного множника у щільностях.

Справді

$$\frac{f_{P_{ij}}(y)}{f_{P_{im}}(y)} = \frac{\sum\limits_{k\neq j}(\lambda_j y+\lambda_k-\lambda_k y)^{-2}\prod\limits_{\substack{u=1,\\ u\neq k,j}}^{N}(\lambda_j y+\lambda_u-\lambda_u y)^{-1}}{\sum\limits_{k\neq j}(\lambda_m y+\lambda_k-\lambda_k y)^{-2}\prod\limits_{\substack{u=1,\\ u\neq k,m}}^{N}(\lambda_m y+\lambda_u-\lambda_u y)^{-1}} =$$

$$= \frac{\sum\limits_{k\neq j}(t\lambda_j y+t\lambda_k-t\lambda_k y)^{-2}\prod\limits_{\substack{u=1,\\ u\neq k,j}}^{N}(t\lambda_j y+t\lambda_u-t\lambda_u y)^{-1}}{\sum\limits_{k\neq j}(t\lambda_m y+t\lambda_k-t\lambda_k y)^{-2}\prod\limits_{\substack{u=1,\\ u\neq k,m}}^{N}(t\lambda_m y+t\lambda_u-t\lambda_u y)^{-1}}.$$

Таким чином, параметри $(\lambda_1,\lambda_2,\ldots,\lambda_N)$ будуть інваріантними відносно множення на той самий множник. Отже, можна завжди припускати, що $\hat{\lambda}_{i1}=1$.

Крім того, побудована в лемі 2 щільність дає змогу як оцінювати параметри моделі для кожного конкретного вузла вебпростору зокрема, так і перевіряти умову однорідності переходів, яка була використана як припущення у лемі 1. Для оцінювання параметрів моделі можна користуватися методом максимальної правдоподібності для розподілу (3), при цьому функція правдоподібності для деякого стану буде мати вигляд

$$L(\lambda)=L(\lambda_1,\lambda_2,\ldots,\lambda_N)=$$

$$\prod_{j=1}^{N}\prod_{t=1}^{T}\Lambda(1-y_{tj})^{N-2}\sum_{k\neq j}(\lambda_j y_{tj}+\lambda_k-\lambda_k y_{tj})^{-2}\prod_{\substack{u=1,\\ u\neq k,j}}^{N}(\lambda_j y_{tj}+\lambda_u-\lambda_u y_{tj})^{-1}$$

або логарифмічної функції правдоподібності

$$l(\lambda)=NT\ln(\Lambda)+$$

$$+\sum_{j=1}^{N}\sum_{t=1}^{T}\left[\ln\left(\begin{array}{c}(N-2)\ln(1-y_{tj})\\ \sum\limits_{k\neq j}(\lambda_j y_{tj}+\lambda_k-\lambda_k y_{tj})^{-2}\prod\limits_{\substack{u=1,\\ u\neq k,j}}^{N}(\lambda_j y_{tj}+\lambda_u-\lambda_u y_{tj})^{-1}\end{array}\right)\right],$$

де $y_t=(y_{t1},\ldots,y_{tN})$ — вектор спостережень (пропорцій) переходу зі стану i у стани $1,2,\ldots,N$, час спостережень змінюється від 1 до T.

У [7, 8] розглянуто методи вибору оптимальної кількості кластерів, проте використання цих методів не дає змоги зробити висновки про належність вебсторінок одному чи різним кластерам. Оскільки однорідність вибірок буде перевірятися на різних вибірках по різних вебсторінках, для визначення належності одному кластеру двох вебсторінок можна використати три підходи.

Підхід 1. Використання тесту Колмогорова–Смірнова. Недоліком цього підходу є те, що тест є непараметричним з невисокою потужністю.

Підхід 2. Побудова інтервалів надійності із використанням інформації за Фішером. Такий підхід ґрунтується на застосуванні статистичних методів із більш високою потужністю, зокрема, на можливості використання асимптотичних властивостей оцінок параметрів розподілу та побудови інтервалів надійності з використанням інформації за Фішером [9, 10].

Загальний алгоритм перевірки гіпотези про належність *i*-го та *j*-го станів одному кластеру (параметри для двох станів: $\lambda^{(i)}$, $\lambda^{(j)} \in R_+^N$, визначаються за допомогою гіпотези H_0: $\lambda^{(i)} = \lambda^{(j)}$, на відміну від альтернативної гіпотези H_A: $\lambda^{(i)} \neq \lambda^{(j)}$ щодо параметрів розподілу двох станів із сімейства (4)), можна сформулювати так.

Крок 1. Оцінка $\hat{\lambda}^{(i)}$ параметрів для *i*-го стану.

Крок 2. Обчислення інформації за Фішером $I_T(\hat{\lambda}^{(i)})$ для розподілу (4) та побудова інтервалу надійності $CI_\alpha^{(i)}$ (α — рівень значущості) для розподілу $\boldsymbol{N}(\hat{\lambda}^{(i)}; I_T^{-1}(\hat{\lambda}^{(i)}))$ для першої вибірки.

Крок 3. Оцінка $\hat{\lambda}^{(j)}$ параметрів для *j*-го стану.

Крок 4. Перевірка належності оцінки $\hat{\lambda}^{(j)}$ інтервалу надійності $CI_\alpha^{(i)}$ та висновок про належність *i*-го та *j*-го станів одному кластеру.

Підхід 3. **Використання bootstrap-методу моделювання інтервалу надійності $\boldsymbol{CI_\alpha^{(i)}}$ [11, 12].** Згідно з цим методом інтервал надійності $CI_\alpha^{(i)}$ будується так: на основі оцінки $\hat{\lambda}^{(i)}$ генеруються вибірки $y_t^{(i)} = (y_{t1}^{(i)}, \ldots, y_{tN}^{(i)})$, $t = 1, \ldots, B$ (B — кількість ітерацій у bootstrap-методі); за вибірками $y_t^{(i)}$ здійснюється оцінювання параметра λ для розподілу (4); на основі отриманих оцінок будується інтервал надійності $CI_\alpha^{(i)}$.

У [7, 8] розроблено спектральний метод оцінювання оптимальної кількості кластерів у задачах кластеризації на графах, який задається за допомогою матриці суміжності. У цих роботах не розглядаються задачі визначення належності одному кластеру декількох станів, що звичайно є важливою прикладною проблемою. Більшість класичних робіт стосується гауссівських ансамблів, для яких робиться припущення про нормальний розподіл елементів матриці та незалежність елементів у тому чи іншому сенсі.

У запропонованій роботі зроблено припущення про експоненціальний розподіл елементів матриці суміжності, що дає змогу точніше описувати характеристики відповідного графу. Крім того, розглянуто випадок залежних елементів матриці, що дещо ускладнює оцінювання асимптотичних властивостей матриці. Основні результати роботи сформульовані у лемі 1 і лемі 2 та стосуються розподілу елементів матриці переходу (2), побудованого на матриці суміжності графу (1). За умови належності всіх вершин графу одному розподілу показників розподіл елементів матриці переходу P не буде залежати від показника λ, що може бути розглянуто як критерій належності одному кластеру певної множини вершин. Також сформульовано три підходи перевірки гіпотез про входження різних вершин до одного кластера, які базуються на оцінці інформації за Фішером та на bootstrap-методі.

У наступних роботах планується зробити перевірку відповідних гіпотез на реальних даних.

СПИСОК ЛІТЕРАТУРИ

1. Tao T. Topics in Random Matrix Theory. *American Mathematical Society*, 2023. 282 p.
2. Chen Y., Sanghavi S., Xu H. Improved graph clustering. *IEEE Transactions on Information Theory*. 2014. Vol. 60, N 10. P. 6440–6455. https://doi.org/10.1109/TIT.2014.2346205.
3. Schaeffer S.E. Graph clustering. *Computer Science Review.* 2007. Vol. 1, Iss. 1. P. 27–64. https://doi.org/10.1016/j.cosrev.2007.05.001.
4. Tsitsulin A., Palowitch J., Perozzi B., Müller E. Graph clustering with graph neural networks. *Journal of Machine Learning Research*. 2023. Vol. 24. P. 1–21.
5. Brandes U., Gaertler M., Wagner D. Experiments on graph clustering algorithms. Di Battista G., Zwick U. (Eds.). Algorithms – ESA 2003. *Lecture Notes in Computer Scienc*e. 2003. Vol. 2832. P. 568–579. https://doi.org/10.1007/978-3-540-39658-1_52.
6. Kondruk N.E., Malyar M.M. Analysis of cluster structures by different similarity measures. *Cybernetics and Systems Analysis*. 2021. Vol. 57, N 3. P. 436–441. https://doi.org/10.1007/s10559-021-00368-4.
7. Kyrychenko O.L., Malyk I.V., Ostapov S.E. Stochastic models in artificial intelligence development. *Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics.* 2021. N 2. P. 53–57. https://doi.org/10.17721/1812-5409.2021/2.7.
8. Kyrychenko O.L., Malyk I.V., Ostapov S.E. Cluster structure analysis of internet networks based on random matrixes. *The International Scientific and Technical Journal «Problems of control and informatics»*. 2022. N 2. P. 37–46. https://doi.org/10.34229/1028-0979-2022-1-4.
9. Robert C.P. The Bayesian Choice. New York: Springer, 2007. 602 p.
10. Taylor S. Clustering financial return distributions using the Fisher information metric. *Entropy*. 2019. Vol. 21, N 2. P. 1–16. https://doi.org/10.3390/e21020110.
11. Chernozhukov V., Chetverikov D., Kato K., Koike Y. High-Dimensional Data Bootstrap. *Annual Review of Statistics and Its Application*. 2023. Vol. 10. P. 427–449. https://doi.org/10.1146/annurev-statistics-040120-022239.
12. Sulafah M. Salem Binhimd, Zakiah I. Kalantan. Bootstrap approach for clustering method with applications. *International Journal of Advanced and Applied Sciences*. 2023. Vol. 10, Iss. 3. P. 189–195. https://doi.org/10.21833/ijaas.2023.03.023.

O.L. Kyrychenko

A CLASS OF RANDOM MATRICES

Abstract. The paper examines methods for assessing the distribution of elements in a stochastic matrix assuming an exponential distribution of elements in the corresponding adjacency matrix of a graph. Two cases are considered: the first assumes homogeneity of all graph vertices, while the second assumes heterogeneity in the distribution of vertices with corresponding density calculations. Hypothesis testing tests are formulated for the respective distributions to determine the membership of two graph vertices in the same cluster.

Keywords: random matrix, artificial intelligence, discrete Markov chain, bootstrap method, Fisher information.

Надійшла до редакції 22.06.2023