Оценивание сходства разбиений графов на пересекающиеся сообщества

Марина Ананьева[1], Анвар Курмуков[1,2], Юлия Додонова[2], Леонид Жуков[1], Борис Гутман[3], Джошуа Фасковитс[3], Неда Джаханшад[3], Пол Томпсон[3]

[1] Национальный исследовательский университет «Высшая школа экономики», Россия

[2] Институт проблем передачи информации им. А.А. Харкевича РАН, Россия

[3] Центр обработки генетических изображений, Институт нейровизуализации и информатики Стивенсона, Университет Южной Калифорнии, США

Аннотация В данной работе мы анализируем структуру сообществ коннектомов – графов, построенных на основе данных нейровизуализации мозга человека. Мы рассматриваем три подхода к определению кластерных структур на коннектомах: непересекающиеся, пересекающиеся и нечеткие сообщества. Для каждого из подходов мы анализируем соответствующую метрику оценивания схожести разбиений на коннектомах: индекс Рэнда и его обобщения для пересекающихся сообществ – Омега индекс и нечеткий индекс Рэнда, различающиеся по допущению степени принадлежности вершины к нескольким сообществам одновременно. Мы проводим сравнительный анализ этих подходов, оценивая их с точки зрения информативности для последующей классификации, на примере пациентов с болезнью Альцгеймера и без этой патологии. В качестве классификаторов мы используем метод k-ближайших соседей и метод опорных векторов. Наилучшие результаты в задаче классификации были получены при рассмотрении пересекающихся сообществ с использованием метода опорных векторов (ROC AUC 0.83).

Keywords: пересекающиеся сообщества; метрики оценки расстояний; машинное обучение на графах

1 Введение

В области нейронаук коннектомами называют графы, построенные на основе данных нейровизуализациии мозга человека (МРТ снимков головного мозга). В таких графах в качестве вершин заданы различные области коры головного мозга, и ребра представляют связи между ними. Графы являются неориентированными и взвешенными: наличие и сила связи определяется по числу трактов (пучков белого вещества), выявленных с помощью магнитно-резонансной томографии. Так как вершинами графов служат области коры головного мозга, то все различия между графами заключены в ребрах и весах ребер.

В современных исследованиях встречается множество попыток решать различные задачи классификации на основе коннектомов. Например, в работе [2] – гендерная классификация, в [8] продемонстрировано, что графы, восстановленные по различным снимкам одного и того же человека, с высокой точностью отличаются от коннектомов разных людей. Также встречаются задачи фенотипной классификации, например, различение здоровых людей от страдающих расстройствами аутистического спектра [5], шизофренией [11], [10], с болезнью Паркинсона и без данной патологии [3].

В работе [5] был предложен оригинальный алгоритм решения задачи классификации коннектомов, основанный на сравнений кластерной структуры графов: расстояние между двумя графами измеряется как расстояние между разбиениями этих графов на сообщества. В работе [6] эта идея была развита на случай пересекающихся сообществ. В настоящей работе мы предлагаем рассмотреть наиболее общий случай кластерной структуры и демонстрируем работу алгоритма на примере задачи классификации больных с болезнью Альцгеймера, умеренными когнитивными нарушениями и здоровыми людьми.

2 Кластерная структура коннектомов

2.1 Виды кластеров

В литературе выделяют три типа кластеров: непересекающиеся кластеры или разбиения (non-overlapping communities or partitions); пересекающиеся кластеры (overlapping crisp communities), нечеткие кластеры (overlapping fuzzy communities) см. Рис. 1.

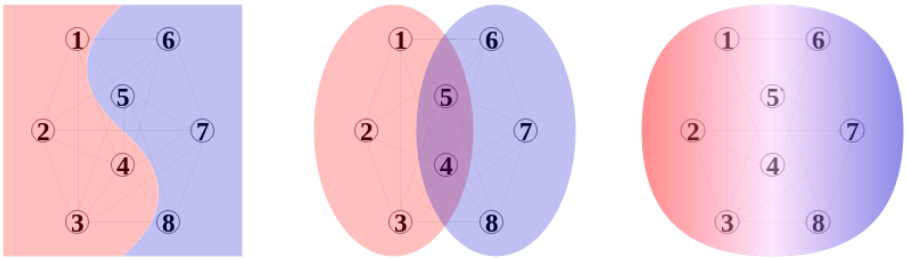

Рис. 1. Пример применения различных типов кластеризации к одному и тому же набору данных. Слева направо: непересекающиеся кластеры, пересекающиеся кластеры, нечеткие кластеры.

Непересекающимися кластерами или разбиениями множества $X = \{x_1, x_2, \dots x_n\}$ называют множество $U = \{U_1, U_2, \dots U_l\}$, такое, что $\bigcup_i U_i = X$, $U_i \cap U_j =$

$\emptyset$, $\forall\ i \neq j \in \{1, \ldots, l\}$. Для хранения такой кластерной структуры часто используют следующий подход: хранится вектор принадлежности p, длины n (для левой картинки на Рис. 1 таким вектором будет $[1, 1, 1, 1, 0, 0, 0, 0]$), в котором i-ый элемент принимает значение метки класса, а набор меток состоит из l уникальных чисел (например от 1 до l). Таким образом, каждому объекту сопоставлена метка кластера. Другим возможным способом будет использование матрицы P размера $n \times l$, в которой i-ому элементу множества X соответствует строка длины l, и j-ый элемент этой строки равен 1 в том случае, если $x_i \in U_j$ (причем в каждой строке не более одного не нулевого элемента).

Пересекающиеся кластеры возникают в случае, когда один и тот же объект принадлежит разным кластерам, но степень принадлежности к этим кластерам не определена. К такому виду относится центральный случай на Рис. 1. Вершины 4 и 5 принадлежат как красному, так и синему кластерам. Такая структура пересекающихся сообществ может быть отражена с использованием упомянутой выше матрицы P размера $n \times l$, в строках которой теперь может быть более одного ненулевого элемента.

Нечеткие кластеры являются самым общим случаем кластеров. Каждый объект не только может принадлежать разным кластерам одновременно, но также известна его сила принадлежности к каждому из них. Например, на правом изображении Рис. 1 вершину 6 можно считать принадлежащей синему кластеру с силой 0.9, а красному – с силой 0.1. Вершины 4 и 5 с равной силой 0.5 принадлежат как синему, так и красному кластеру. Такая структура также может храниться с помощью матрицы P размера $n \times l$, где n - число объектов множества, а l - число кластеров, но теперь элементами этой матрицы могут быть не только значения 0 и 1, но и любые неотрицательные числа (отнормировав которые на сумму по строкам матрицы P мы всегда можем получить матрицу с элементами, принадлежащими отрезку $[0, 1]$). Для этого мы пользовались неотрицательным матричным разложением.

2.2 Неотрицательное матричное разложение

Для нахождения кластерной структуры коннектомов использовался подход, основанный на неотрицательном матричном разложении специального вида: разложение неотрицательной матрицы $A \in \mathbb{R}_+^{n \times n}$ в произведение неотрицательных матриц $W \in \mathbb{R}_+^{n \times k}$ и $H \in \mathbb{R}_+^{k \times n}$:

$$\begin{aligned} \| A - WH \|_2 &\to \min, \\ W \geq 0,\ H &\geq 0. \end{aligned} \tag{1}$$

В нашем случае, в качестве A была выбрана бинаризованная матрица смежности каждого коннектома. Так как элементы матриц W и H – неотри-

цательны, то матрицу H^T размера $n \times k$ можно интерпетировать как матрицу **нечеткой кластерной структуры** n объектов (вершин коннектома) на k нечетких кластеров:

$$P_{fuzzy} = H^T. \tag{2}$$

После того как были получены **нечеткие кластеры** P_{fuzzy}, от них легко перейти к **пересекающимся кластерам** P_{crisp}. Для этого достаточно отнормировать матрицу P_{fuzzy} на сумму значений по строкам, отбросить значения матрицы, не превышающие некоторый порог δ, и приравнять все оставшиеся значения к единице:

$$P^{ij}_{crisp} = \begin{cases} 1, \ if \ \ P^{ij}_{fuzzy} \geq \delta \\ 0, \ if \ \ P^{ij}_{fuzzy} < \delta \end{cases} . \tag{3}$$

Так же просто перейти и к **непересекающимся сообществам**, достаточно в каждой строке матрицы нечетких кластеров оставить максимальный элемент, отбросив остальные:

$$P^{ij}_{partition} = \begin{cases} 1, \ if \ \ P^{ij}_{fuzzy} = max_j(P^i_{fuzzy}) \\ 0, \ otherwise \end{cases} . \tag{4}$$

2.3 Способы сравнения кластерных структур

В наших экспериментах в зависимости от вида кластерной структуры (пересекающиеся, непересекающиеся и нечеткие сообщества) мы использовали три меры для оценки сходства разбиений графов: Индекс Рэнда (Rand Index) [9], Омега Индекс (Omega Index)[1], Нечеткий Индекс Рэнда (Fuzzy Rand Index)[4]. Данные меры сходства принимают максимальное значение 1, если две кластерные структуры одинаковы, и близкие к нулю значения в обратном случае.

Индекс Рэнда является одним из первых предложенных методов для измерения сходства двух кластерных структур (в случае непересекающихся кластеров). Он рассчитывается как отношение числа согласованных пар объектов в двух разбиениях к числу всех возможных пар. Пусть дано множество X состоящее из n элементов: $X = \{x_1, x_2, \dots x_n\}$. Рассмотрим два различных разбиения этого множества на сообщества: $U = \{U_1, U_2, \dots U_l\}$ и $V = \{V_1, V_2, \dots V_k\}$, где l и k - число кластеров в разбиениях U и V соответственно. Для их сравнения воспользуемся формулой:

$$RI(U, V) = \frac{a + b}{a + b + c + d}, \tag{5}$$

a - количество пар $\{x_i, x_j\}$, таких что $\{x_i, x_j\} \subseteq V_t$ и $\{x_i, x_j\} \subseteq U_r$;
b - количество пар $\{x_i, x_j\}$, таких что $x_i \in V_{t_1}, x_j \in V_{t_2}, x_i \in U_{r_1}, x_j \in U_{r_2}$;
c - количество пар $\{x_i, x_j\}$, таких что $\{x_i, x_j\} \subseteq V_t$, но $x_i \in U_{r_1}, x_j \in U_{r_2}$;

d - количество пар $\{x_i, x_j\}$, таких что $x_i \in V_{t_1}, x_j \in V_{t_2}$ и $\{x_i, x_j\} \subseteq U_r$.

При этом, $i \neq j$, $t_1 \neq t_2$, $r_1 \neq r_2$, $1 \leq i, j \leq n$, $1 \leq t, t_1, t_2 \leq l$, $1 \leq r, r_1, r_2 \leq k$. Таким образом, $(a + b)$ - это количество пар объектов, которые в обоих разбиениях находятся либо в одном кластере, либо в разных. В то время, как $(c+d)$ - количество пар объектов, которые в одном разбиении находятся в одном кластере, а в другом разбиении находятся в разных.

Омега Индекс - это некоторое обобщение Индекса Рэнда для случая пересекающихся сообществ [1]. Данный индекс измеряет общее число согласованных пар вершин в двух покрытиях графов, где согласованными называются вершины, принадлежащие одинаковому количеству сообществ. Таким образом, значение Омега Индекса зависит от количества пар вершин, принадлежащих одному, двум или большему числу кластеров. **Нечеткий Индекс Рэнда**, предложенный Е. Хюллермейером и М. Рифки [4], так же является обобщением Индекса Рэнда, но для кластеров с нечеткой структурой.

Все три меры имеют под собой одну теоретико-множественную природу и эквивалентны для случая непересекающихся сообществ. Кроме того, на данный момент Нечеткий Индекс Рэнда является единственным существующим способом измерять сходство между двумя нечеткими кластерными структурами.

3 Классификация коннектомов на основании структуры сообществ

Для того, чтобы определить, какой из трех подходов к определению расстояния между графами коннектомов оказывается наболее информативен, мы будем решать несколько бинарных задач классификации, используя расстояния между непересекающимися, пересекающимися или нечеткими сообществами. Подход, с помощью которого будет достигнут наивысший результат, будем считать наилучшим. Сформулируем задачу более формально: пусть дан набор графов $\mathbb{G} : \{(G_1, y_1), (G_2, y_2), \dots (G_m, y_m)\}$. Необходимо построить модель классификации, обучающуюся на части выборки $\mathbb{G}$ и предсказывающую метки классов оставшейся части выборки. Сначала найдем сообщества каждого графа и построим матрицы попарных расстояний:

1. Для каждого из m графов G найти его сообщества:
   (a) Пересекающиеся, используя формулу 3 (P_{crisp});
   (b) Непересекающиеся, используя формулу 4 ($P_{partition}$);
   (c) Нечеткие, используя NMF (P_{fuzzy}).
2. Посчитать матрицы попарных расстояний, используя полученные сообщества:
   (a) $K_{ij}^{Omega} = 1 - Omega(P_{Omega}^{i}, P_{Omega}^{j})$, $i, j = 1 \dots m$;
   (b) $K_{ij}^{RI} = 1 - RI(P_{partition}^{i}, P_{partition}^{j})$, $i, j = 1 \dots m$;
   (c) $K_{ij}^{fRI} = 1 - fRI(P_{fuzzy}^{i}, P_{fuzzy}^{j})$, $i, j = 1 \dots m$.

Полученные матрицы $K_{ij}^{Omega}, K_{ij}^{RI}, K_{ij}^{fRI}$ можно подавать на вход классификатора. Воспользуемся методом k-ближайших соседей и методом опорных векторов, для последнего предварительно используя экспоненцирование матрицы попарных расстояний:

$$K_{SVM} = e^{-\alpha K}. \tag{6}$$

4 Результаты

4.1 Набор данных

В данной работе использовался набор данных, подготовленный в рамках Alzheimer‘s Disease Neuroimaging Initiative. Данные собраны по 228 участникам (756 снимков). Средний возраст пациентов на первичном осмотре 72.9 ± 7.4 лет (96 женщин и 132 мужчин). Для каждого человека имеется не менее 1 и не более 6 снимков МРТ головного мозга. Данные включают 47 пациентов с диагностированной врачами болезнью Альцгеймера (136 снимков AD), 40 пациентов с ранними нарушения когнитивных функций (147 снимков LMCI), 80 пациентов с поздней стадией умеренного нарушения когнитивных функций (238 сканов по EMCI) и 61 пациент без патологии (190 сканов NC). Для разметки вершин графа был использован атлас Десикана-Киллиани (DK) [7], который включает 68 областей головного мозга. Веса ребер в исходной матрице кортикальных связей пропорциональны количеству трактов, обнаруженных алгоритмом [12].

4.2 Результаты экспериментов

В работе решались четыре бинарные задачи классификации: диагноз болезни Альцгеймера против нормы (AD vs NC), диагноз болезни Альцгеймера против поздних когнитивных нарушений (AD vs LMCI), поздние когнитивные нарушения против ранних когнитивных нарушений (LMCI vs EMCI), ранние когнитивные нарушения против нормы (EMCI vs NC). Результаты работы SVM классификатора на всех построенных ядрах и k-ближайших соседей по 5-fold кросс-валидации с усреднением по 50 различным разбиениям представлены в Таблицах 1 и 2. При проведении кросс-валидации учитывалось, чтобы часть графов относилась к одному и тому же человеку.В Таблице 1 показаны результаты решения задачи методом k-ближайших соседей для трех разных кластерных структур - непересекающихся (Partition), пересекающихся (Crisp) и нечетких (Fuzzy) сообществ. Как можно видеть, наилучшее качество, оцененное по ROC кривой, достигается при сравнении крайних групп: пациентов, у которых врачи диагностировали болезнь Альцгеймера, с участниками исследования без патологии. Значимо хуже результаты получены для сравнения пациентов с пограничными состояниями при попытке разграничить между собой поздние и ранние когнитивные нарушения, а также ранние нарушение против нормы. В то же время, при

сравнении результатов по каждой из четырех групп по трем кластерным структурам, можно выделить схожие результаты. Тем не менее, подход с пересекающимися сообществами показывает более высокие результаты (до $.776 \pm .013$ по ROC AUC), чем непересекающиеся сообщества (до $.735 \pm .012$ по ROC AUC).

Таблица 1. KNN. Результаты решения четырех задач классификации (слева направо), сравнение трех ядер (сверху вниз) с использованием метода ближайшего соседа. Все значения даны в формате среднее ± стандартное отклонение по ROC AUC.

Task	AD vs NC	AD vs LMCI	LMCI vs EMCI	EMCI vs NC
Partition	$.735 \pm .012$	$.722 \pm .015$	$.499 \pm .018$	$.597 \pm .014$
Crisp	$.776 \pm .013$	$.714 \pm .012$	$.514 \pm .013$	$.550 \pm .015$
Fuzzy	$.768 \pm .015$	$.705 \pm .018$	$.555 \pm .014$	$.600 \pm .014$

Полученные результаты с использованием метода опорных векторов (SVM) в Таблице 2 значимо лучше, чем при методе ближайших соседей. Наилучшее качество модели получено при сравнении пациентов с болезнью Альцгеймера против группы нормы при пересекающихся кластерах ($.831 \pm .009$). Как мы видим, для сравнения крайних, наиболее раличающихся между собой групп нормы и болезни Альцгеймера, наилучший результат показывает кластерная структура с пересекающими сообществами с четко определенной принадлежностью вершин (crisp).

Таблица 2. SVM. Результаты решения четырех задач классификации (слева направо), сравнение трех ядер (сверху вниз) с использованием метода опорных векторов. Все значения даны в формате среднее ± стандартное отклонение по ROC-AUC.

Task	AD vs NC	AD vs LMCI	LMCI vs EMCI	EMCI vs NC
Partition	$.795 \pm .010$	$.762 \pm .018$	$.523 \pm .028$	$.628 \pm .018$
Crisp	$.831 \pm .009$	$.725 \pm .027$	$.528 \pm .029$	$.591 \pm .022$
Fuzzy	$.786 \pm .006$	$.703 \pm .016$	$.542 \pm .018$	$.616 \pm .015$

5 Заключение

В данном исследовании мы решали задачу бинарной классификации пациентов, применяя четыре попарные сравнения групп пациентов с болезнью Альцгеймера, переходной стадией или без патологии на графах, отражающих структуру головного мозга человека - коннектомах. Мы предположили, что можно различать фенотипы пациентов на основании схожести разбиения графов внутри одной группы и различий с другими, и на основании этой гипотезы протестировали три различные метрики по оценке схожести разбиений графов и два метода классификации. В качестве способа кластеризации выбор был сделан в пользу неотрицательного матричного разложения (NMF), применяя которую можно легко перейти к рассмотрению пересекающихся сообществ. Каждый граф был разбит на оптимальные подграфы, чтобы далее получить вектора меток принадлежности вершин к подграфам в разбиениях. необходимо было оценить попарные расстояния между разбиениями, рассматривая их в качестве расстояний между коннектомами. На основе полученных матриц из попарных расстояний были построены ядра для классификатора. Всего было рассмотрено три различных типа кластерных структур: пересекающиеся, непересекающиейся и нечеткие сообщества. Для каждого из двух использованных алгоритмов кластеризации - k-ближайших соседей и метода опорных векторов - три метрики по оценке схожести разбиений оказали разное влияние на качество классификации, сравнение которых показало преимущество применения непересекающихся сообществ.

По полученным результатам можно сказать, что графы, построенные по модульной сетевой структуре головного мозга, могут достаточно хорошо захватывать важную информацию об индивидуальных особенностях связей кортикальных областей мозга, которая может быть использована для определения и классификации фенотипов пациентов с нейродегеративными заболеваниями. Рассмотренные метрики схожести разбиений (Индекс Рэнда, Омега Индекс, Нечеткий Индекс Рэнда) выдают относительно похожие высокие и относительно схожие результаты по ROC-AUC, что позволяет говорить о применимости данных метрик для решения задачи классификации на коннектомах. Наилучшие результаты были достигнуты при рассмотрении пересекающихся сообществ и применении метода опорных векторов SVM для крайних групп пациентов с болезнью Альцгеймера и без патологии (ROC AUC $.831 \pm .009$).

Дальнейшие исследования могут быть направлены на тестирование работы алгоритма на других наборах данных, чтобы сравнить устойчивость полученных результатов.

Благодарности. Данные, использованные в исследовании, были собраны в рамках Инициативы по нейровизуализации болезни Альцгеймера. Полный список исследователей и документации доступен по ссылке `adni.loni.usc.edu`.

Исследование проведено в Институте проблем передачи информации им. А.А.Харкевича РАН (секции 2-5) при поддержке Российского научного фонда, проект 17-71-10262.

Список литературы

1. Collins, L.M., Dent, C.W.: Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. Multivariate Behavioral Research 23(2), 231–242 (1988)
2. Dodonova, Y., Petrov, D., Zhukov, L.: Сравнение эффективности ядер svm классификатора для различения пола на основе структурных коннектом. Тр. конф. Информационные технологии и Системы pp. 1155–1167 (2015)
3. Galvis, J., Mezher, A., Ragothaman, A.e.a.: Effects of epi distortion correction pipelines on the connectome in parkinson's disease. SPIE Medical Imaging, 9784(3):97843D (2016)
4. Hullermeier, E., Rifqi, M.: A fuzzy variant of the rand index for comparing clustering structures. Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA-EUSFLAT 2009 pp. 1294–1298 (2009)
5. Kurmukov, A., Dodonova, Y., Zhukov, L.E.: Classification of normal and pathological brain networks based on similarity in graph partitions. Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference, pp. 107–112 (2016)
6. Kurmukov, A., Dodonova, Y., Zhukov, L.E.e.a.: Classifying phenotypes based on the community structure of human brain networks. 20th International Conference on Medical Image Computing and Computer Assisted Intervention 2017, September 10-14, Quebec, Canada [accepted] (2017)
7. McDaid, A.F., Greene, D., Hurley, N.: Normalized mutual information to evaluate overlapping community finding algorithms (2011)
8. Petrov, D., Gutman, B., Ivanov, A., Faskowitz, J., Jahanshad, N., Belyaev, M., Thompson, P.: Structural connectome validation using pairwise classification. Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium, pp. 451–455 (2017)
9. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66(336), 846–850 (1971)
10. Salgado-Pineda, P., Junque, C., Vendrell, P.e.a.: Decreased cerebral activation during cpt performance: structural and functional deficits in schizophrenic patients. Neuroimage, vol. 21, pp. 840–847 (2011)
11. Sui, J., Castro, E., Hao, H., Bridwell, D.e.a.: Combination of fmri-smri-eeg data improves discrimination of schizophrenia patients by ensemble feature selection. the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'14) (2014)
12. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, pp. 2837–2854 (2010)