Анализ социальных сетей. Лекция 4
Структурные характеристики сетей

Михаил Пожидаев

21 ноября 2024 г.

Ассортативное смешивание

Связи узлов по назначенным свойствам

Назначим:

каждому узлу некоторое свойство, значение которого может быть качественным (категориальным, дискретным) или количественным (числовым, непрерывным).

Сеть называется ассортативной, если вероятность появления связи выше для узлов, у которых значение некоторого свойства совпадает. Социальные сети чаще всего ассортативны по возрасту или региону проживания пользователя.

Корреляция

Для количественного случая ассортативность сети можно рассматривать как корреляцию свойства на разных концах рёбер.

Модулярность

Пусть задано m групп $g_1, g_2, \ldots, g_m$ и все узлы из N принадлежат к одной из них. Примем $\delta_{ij} = 1$, если $g_i = g_j$ и ноль в противном случае. Оценка ассортативности выражается как:

$$q_1 = \frac{1}{2} \sum_{i,j \in N} (A_{ij} \delta_{ij})$$

Средняя ассортативность

Получение оценочной характеристики

Полезно сравнивать ассортативность со средней ассортативностью:

$$q_2 = \frac{1}{2} \sum_{i,j \in N} (p_{ij} \delta_{ij}),$$

где p_{ij} — вероятность существования ребра между i и j. Удобно нормировать значения на количество рёбер m.

Модулярность сети

$$q = \frac{q_1 - q_2}{m}$$

Структурное подобие

Поиск вероятных знакомых

Мера структурного подобия принимается равной 1, если совпадают все смежные узлы, и 0, если все смежные узлы различаются.

Косинусное расстояние

Определяется как скалярное произведение векторов, делённое на произведение их длин. Векторами принимаются соответствующие строки в матрице смежности.

Вычисляется как мера схожести или сонаправленности двух векторов.

Коэффициент Жаккара

$$\frac{c}{a+b-c},$$

где a — степень первой вершины, b — степень второй вершины, а c — количество общих смежных узлов.

Сообщества

Определение и свойства

Определение

Сообщество — это группа узлов в сети, для которой наблюдается тенденция быть более связанными друг с другом, чем с узлами за пределами этой сети.

Свойства:

- сообщество можно рассматривать как подмножество узлов в сети, стремящихся формировать клику наибольшего размера.
- если сумма связей внутри сообщества превышает сумму связей за пределами сообщества, говорят, что это сильное сообщество.
- в случайной сети сообществ нет.

Иерархическая кластеризация

Выявление сообществ

Очевидно, но недостижимо!

Выделение сообществ можно было бы осуществить перебором узлов с проверками, но это вычислительно недостижимая задача.

Пограничные состояния

Все вершины входят в свои единичные сообщества, и все вершины входят в одно общее сообщество.

В зависимости от выбора начального пограничного состояния производят иерархический поиск сверху путём разделения или снизу путём объединения с выбором оптимальных значений оценок, но что можно считать подобными оценками?

Оценки при кластеризации

Как распознать сообщества?

Восходящий поиск

Выбирается пара текущих сообществ таким образом, чтобы они формировали самое сильное новое сообщество при объединении по всем парам, и объединяются.

Нисходящий поиск

Среди всех текущих сообществ производится попытка разделения таким образом, чтобы получилась наиболее слабая пара, и соответствующее сообщество разделяется.

Если применить понятие ассортативности для случая, когда $\delta_{ij} = 1$ только в том случае, когда узлы i и j входят в одно сообщество, то качество разделения можно определять по значению модулярности q.

Развитие идеи

Бывают расширенные формулировки постановки задачи, при которых допускается вход узла в несколько сообществ или нечёткая принадлежность к сообществу в долях единицы.

Расстояние между сообществами

Три способа оценки

1. *Минимальное расстояние:* используется путь минимальной длины между всеми парами узлов из двух сообществ.
2. *Среднее расстояние:* используется средняя длина пути между парами узлов.
3. *Максимальное расстояние:* используется максимальная длина пути между парами узлов.

Кодирование данных

Битовое представление

Кодирование чисел от 0 до 3:

0=00, 1=01, 2=10, 3=11.

Кодирование чисел от 0 до 7:

0=000, 1=001, 2=010, 3=011, 4=100, 5=101, 6=110, 7=111.

Энтропия

Информационная «полезность» сообщений

Пусть задано множество информационных сообщений $S = \{s_1, s_2, \ldots, s_n\}$, для которых известно математическое ожидание появления каждого из них $p_1, p_2, \ldots, p_n$. Чем меньше вероятность, тем выше информационная полезность сообщения.

Информационная энтропия:

Клод Шеннон показал, что величина неопределённости последовательности сообщений, которую он назвал информационной энтропией, выражается как:

$$H = \sum_{i=1}^{n} (-p_i \log_2 p_i)$$

Мера измерения!

Информационная энтропия измеряется в битах! Чем больше неопределённость, тем больше бит требуется для кодирования.

Алгоритм Infomap

Оценка качества выделения сообществ

Пусть задано разбиение на сообщества с обозначением буквой сообщества, а индексом — члена сообщества ($a_1, a_2, b_1, b_2, \dots$). Дан некоторый случайный путь по сети, условно выраженный последовательностью $a_1, a_2, b_1, b_2, b_3, c_1, \dots$. Вводится битовое кодирование узла внутри сообщества и событий перехода между сообществами. Выход не должен повторять код узлов.

Оптимальное кодирование

Оптимальное кодирование пути — кодирование наименьшей длины. Оптимальное разбиение на сообщества — разбиение с наименьшим оптимальным кодированием.

Вычисление!

Оценить длину оптимального кодирования можно по определению энтропии, поскольку для сети известны математические ожидания появления узла и переходов между сообществами.

Спасибо за внимание!

Всё о курсе: https://marigostra.ru/materials/networks.html
E-mail: msp@luwrain.org
Канал в Телеграм: https://t.me/MarigostraRu