Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет компьютерных наук
Магистерская программа математических методов оптимизации и стохастики

Курсовая работа

Выделение пересекающихся сообществ во взвешенных графах

-| ЧЕРНОВИК |-

Выполнил:
студент группы м15МОС
Славнов Константин Анатольевич

Научный руководитель:
к.ф.-м.н.
Панов Максим Евгеньевич

Москва, 2016

Содержание

1 Введение

-| Общие слова про тему — выделение пересекающихся сообществ.
Актуальность — зачем выделение сообществ, для каких задач надо.
Зачем пересекающиеся, про новизну анализа взвешенных графов.
Про ключевые результаты работы. |-

В данной работе будет рассмотрена задача выделения сообществ — группы вершин в графе, плотно связанных между собой. На текущий момент известно множество подходов и методов для выделения непересекающихся сообществ [1]. Гораздо меньше внимания уделено случаю пересекающихся групп. В данной работе будет предложен новый метод решения задачи в случае взвешенных графов. Метод основан на алгоритме BigClam [2], который создан для не взвешенных графов. Можно сказать, что новый метод является обобщением модели BigClam на взвешенный случай.

Работа начинается с постановки задачи и подробного описания метода BigClam. Особое внимание уделено методу инициализации. Будет показано, как небольшими усилиями можно улучшить предложенный метод инициализации по начальному значению функционала, что ускоряет и немного уточняет итоговый результат.

Далее речь пойдет о подходах обобщения метода на случай взвешенного графа. В начале рассмотрено самое простое и интуитивное обобщение, после чего предложена усложненная модель. Заканчивается работа экспериментами на модельных и реальных данных. К сожалению, реальных взвешенных графов с известным разбиением на пересекающиеся сообщества не удалось найти. Поэтому основные выводы работы будут по экспериментам с модельными данными.

-| Добавить выводы в общих словах, когда они будут сформулированы. |-

2 Постановка задачи

-| Описание общего подхода и принципа. |-

Общий метод базируется на следующем очевидном наблюдении, что чем в большее количество сообществ входят одновременно две вершины, тем больше вероятность, что они будут соединены ребром, что подтверждается на реальных данных [2]. Наша модель должна учитывать этот факт.

Представим, что каждая вершина графа v взаимодействует с сообществом A с некоторой силой F_{vA}. Нулевая сила означает отсутствие взаимодействия. Такую модель можно представить как двусвязный граф, из вершин исходного графа в первой компоненте и вершин-сообществ во второй (рис. 1). Отметим, что подобная концепция позволяет отразить не только идею пересекающихся сообществ, но и вложенных.

Теперь можно определить силу взаимодействия X_{uv} между вершинами u и v, которая будет определять вероятность появления ребра между вершинами: $X_{uv} = F_u \cdot F_v^T$, где F_u — вектор-строка, составленная из F_{uA} — сил взаимодействия вершин с сообществами графа.

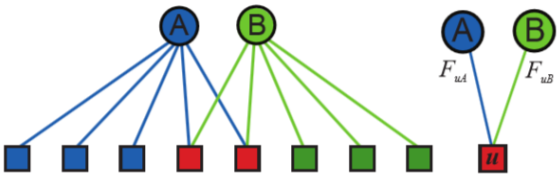

Рис. 1. Двусвязный граф модели BigClam. Сверху вершины-сообщества, снизу вершины исходного графа. Вершины и сообщества взаимодействуют с неотрицательной силой F_{vA}. Ребра, которые соответствуют нулевым весам опущены для большей наглядности.

Таким образом, получено желаемое свойство: чем больше общих сообществ разделяют вершины, тем сильнее они связаны. Определим вероятность появления ребра (u, v) как $p(u, v) = 1 - \exp(-X_{uv})$. Т.е. чем сильнее связаны вершины, тем вероятнее появление ребра между ними. Таким образом получена вероятностная модель. Предполагается, что наблюдаемый нами граф сгенерирован измено из нее.

Итак, кратко опишем Оригинальный метод, его основные предположения.

Основные обозначения. Ниже представлена таблица с основными обозначениями, которые будут использоваться в работе далее.

Обозначение	Описание
$G = (V, E)$	граф
$A \in \mathbb{R}_+^{N \times N}$	матрица смежности
$F \in \mathbb{R}_+^{N \times K}$	матрица силы принадлежности к сообществам
K	количество сообществ
C	множество сообществ

Рассказ про NMF с не квадратичной функцией потерь.

§2.1 Оригинальный метод

Постановка задачи,
предположения,
вывод формул,
схема оптимизации,
примерно как в ноутбуке Math Models.
Про AGM модель и ее релаксацию.

Предположения

1. Каждая вершина v относится к сообществу c с некоторой силой $F_{vc} > 0$.

2. Вероятность появления ребра (u, v), при условии, что вершины u, v находятся в одном сообществе c есть

$$P((u, v)|c) = 1 - \exp(-F_{uc}F_{vc}).$$

3. Каждое сообщество c генерирует ребра независимо друг от друга, а значит, что вероятность появления ребра

$$P(u,v) = 1 - \exp(-\sum_{c \in C} F_{uc} F_{vc}) = 1 - \exp(-F_u F_v^T),$$

$$F = \{F_u\} = \{F_{uc}\} \in \mathbb{R}^{N \times K}.$$

Вероятностная интерпретация

1. Каждые две вершины u, v взаимодействуют с некоторой силой X_{uv}. Чем больше сила, тем больше вероятность $p(u,v)$ появления ребра.

2. Сила взаимодействия вершин u, v определяется сообществами. Каждое сообщество c, в которое входит одновременно 2 вершины дает свой аддитивный вклад в силу их взаимодействия $X_{uv}^{(c)}$.

3. **Предполагаем**, что $X_{uv}^{(c)} \sim \text{Pois}(F_{uc} \cdot F_{vc})$, где $F_{vc} > 0$ сила взаимодействия вершины v и сообщества c. Значит, что

$$X_{uv} \sim \text{Pois}(\sum_c F_{uc} \cdot F_{vc}) = \text{Pois}(F_u \cdot F_v^T).$$

4. **Предполагаем**, что ребро появляется, если $X_{uv} > 0$. Т.е.

$$p(u,v) = \mathbb{P}(X_{uv} > 0) = 1 - \exp(-F_u F_v^T).$$

ε-сообщество Предполагаем, что все вершины относятся к большому ε-сообществу с малой силой ($\approx 10^{-6}$) (т.к. иначе, вершины, не входящие в одно сообщество не могут быть соединены ребром).

Модель Используется метод максимизации правдоподобия.

$$l(F) = log(\mathbb{P}(A|F)) \tag{2.1}$$

$$= \sum_{(u,v) \in E} \log(1 - \exp(-F_u F_v^T)) - \sum_{(u,v) \notin E} F_u F_v^T. \tag{2.2}$$

Схема оптимизации Такая задача является частным случаем NMF (Non-negative matrix factorization): Ищем такую низкоранговую матрицу $F \in \mathbb{R}^{N \times K}$, что она наилучшим образом приближает матрицу A в смысле правдоподобия (на самом деле l_2-норма плохо подходит для восстановления бинарных матриц):

$$\hat{F} = \arg\min_{F \geq 0} D(A, f(FF^T)),$$

$$\text{где} \quad D = -l(F), \quad f(x) = 1 - \exp(-x).$$

Для оптимизации используется блочный координатный спуск с методом проекции градиента на каждом шаге. Фиксируем F_v, оптимизируем по F_u, $u \neq v$. Задача становится выпуклой.

$$\forall u: \quad \arg\max_{F_u \geq 0} l(F_u),$$

$$l(F_u) = \sum_{v \in \mathcal{N}(u)} \log(1 - \exp(-F_u F_v^T)) - \sum_{v \notin \mathcal{N}(u)} F_u F_v^T,$$

где $\mathcal{N}(u)$ — соседи вершины u.

$$\nabla l(F_u) = \sum_{v \in \mathcal{N}(u)} F_u \frac{\exp(-F_u F_v^T)}{1 - \exp(-F_u F_v^T)} - \sum_{v \notin \mathcal{N}(u)} F_v^T.$$

Основная сложность формулы (линейная по размеру графа) во втором слагаемом. Заметим, что

$$\sum_{v \notin \mathcal{N}(u)} F_v^T = \sum_v F_v - F_u - \sum_{v \in \mathcal{N}(u)} F_v.$$

Получаем сложность одной итерации $O(\mathcal{N}(u))$.

Для подбора градиентного шага используем backtracking line search.

Восстановление структуры сообществ Для того, чтобы восстановить исходную структуру сообществ C сравним значение матрицы F с порогом δ. Обозначим за ε вероятность появления ребра в графе (если бы все ребра появлялись равномерно): $\varepsilon = \dfrac{2|V|}{|E| \cdot (|E| - 1)}$. Возьмем δ так, чтобы две вершины принадлежали одному сообществу, если модельная вероятность появления ребра между ними выше чем ε:

$$\varepsilon \leq 1 - \exp(-\delta^2)$$

$$\delta = \sqrt{-\log(1 - \varepsilon)}$$

§2.2 Инициализация

Про Conductance и инициализацию в BigClam.
про ее не совершенство. Идеи по улучшению. Раз, два, три. Тестирование на модельных данных. Тестирование на реальных данных.
Выводы.

3 Новые модели для взвешенных графов.

Наивный переход к взвешенному варианту (деление на вес ребра).
Что-то рассказать про него (?).

Самое простое изменение BigCLAM для обработки взвешенных ребер:

$$l(F) = \sum_{(u,v)\in E} \log(1 - \exp\left(-\frac{F_u F_v^T}{w_{uv}}\right) - \sum_{(u,v)\notin E} \frac{F_u F_v^T}{w_{uv}}.$$

Тем самым, мы получаем, что чем больше вес w_{uv}, тем больше должно быть значение сил F_u и F_v, которые его объясняют.

Т.е. Предположение, что вероятность появления ребра (u, v), при условии, что вершины u, v находятся в одном сообществе c есть

$$P((u, v)|c) = 1 - \exp\left(-F_{uc}F_{vc}\right).$$

Заменяется на предположение, что вероятность появления ребра (u, v) **с весом w_{uv}**, при условии, что вершины u, v находятся в одном сообществе c есть

$$P((u, v)|c, w_{uv}) = 1 - \exp\left(-\frac{F_{uc}F_{vc}}{w_{uv}}\right).$$

Красивая вероятностная интерпретация:

$$X_{uv}^{(c)} \sim \text{Pois}\left(\frac{F_{uc} \cdot F_{vc}}{w_{uv}}\right).$$

Тесты подтверждают работоспособность модели.

§3.1 Gamma Модель

Переход к Гамма моделям. Первоначальный вариант.
Проблема разреженных данных. Очень долгая сходимость.
Переход к Разреженной Гамма модели.

Если мы посмотрим на первоначальную модель, то увидим, что в ней есть скрытые переменные X_{uv}, которые распределены по Пуассону. Данное распределение является дискретным, а веса на ребрах – непрерывные.

Самое главное свойство, которые использовались в выводе — мультипликативность. Так что в качестве непрерывного аналога распределения Пуассона можно взять Гамма распределение, сумма которых (с одинаковым коэффициентом масштаба) не выводит из класса.

Предположения * Вероятность появления ребра с весом w_{uv}, при условии, что вершины принадлежат сообществу c

$$p\left(w_{uv}|c\right) \sim \Gamma\left(k = F_u F_v^T + 1, \theta = 1\right).$$

* Каждое сообщество c генерирует ребра независимо друг от друга, а значит, что вес ребра

$$w_{uv} = \sum_c w_{uv}^c \sim \Gamma\left(\sum_c F_{uc}F_{vc} + 1, 1\right) = \Gamma\left(F_u F_v^T + 1, 1\right).$$

Берем $+1$ для того, чтобы не сталкиваться с распределениями с бесконечной плотностью в нуле. В вероятностном смысле это безусловная (от сообществ) вероятность появления ребра в графе.

Для простоты везде далее будем опускать параметр θ.
Обозначим $K_{uv} = F_u F_v^T + 1$.

Модель

$$l(F) = log\left(\mathbb{P}(A|F)\right) = \sum_{w_{uv}} \log p(w_{uv}) \tag{3.1}$$

$$= \sum_{w_{uv}} \left[-\log\Gamma(K_{uv}) - K_{uv}\log\theta + (K_{uv}-1)\cdot\log w_{uv} - \frac{w_{uv}}{\theta} \right] \tag{3.2}$$

$$= \sum_{w_{uv}} \left[-\log\Gamma(K_{uv}) + (K_{uv}-1)\cdot\log w_{uv} - w_{uv} \right] \tag{3.3}$$

$$= \sum_{w_{uv}} \left[-\log\Gamma\left(F_u F_v^T + 1\right) + F_u F_v^T \cdot \log w_{uv} - w_{uv} \right] \to \max_{F \geq 0}. \tag{3.4}$$

$$\tag{3.5}$$

§3.2 Разреженная Gamma Модель

Введение разреженной составляющей.
Вывод формул. Анализ: Композиция оригинального метода и Гамма модели.
Игрушечные примеры.

4 Про функционалы качества

Про случай пересекающихся сообществ — немного специфики.

§4.1 Модулярность

Модулярность

§4.2 Conductance

Conductance

§4.3 NMI

NMI

5 Данные

§5.1 Модельные Данные

Про модель генерации пару слов.

§5.2 Реальные Данные

Рассказ про то, что мало таких данных с истинным разбиением на сообщества.
Метрики Модулярность и Conductance.
Данные раз.
Данные два.
Данные три.

6 Эксперименты

Общее описание.
С какими методами сравниваемся еще.

§6.1 Эксперименты на модельных данных

Описание. Выводы.

§6.2 Эксперименты на реальных данных

Описание. Выводы.

7 Результаты работы

Выводы.
Что нового сделано.
Направление дальнейших разработок.

8 Список литературы

[1] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
http://www.arxiv.org/abs/0906.0612.

[2] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM, 2013.