Министерство образоваия и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Московский физико-технический институт
(государственный университет)»

Факультет управления и прикладной математики

Кафедра проблем передачи информации и анализа данных

ИССЛЕДОВАНИЕ СЛУЧАЙНЫХ ГЕОМЕТРИЧЕСКИХ ПОЛЕЙ ПОСРЕДСТВОМ ГРАФОВЫХ МЕТРИК

Выпускная квалификационная работа
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:
студент 177 группы _____ Виденеева Анастасия Сергеевна

Научный руководитель:
д. ф.-м. н., старший научный сотрудник _____ Чеботарев Павел Юрьевич

Москва 2015

Оглавление

Введение

Во многих задачах машинного обучения графы используются для моделирования связей между объектами. Например, анализ социальных графов и сетей, создание рекомендательных систем, транспортные задачи.

Наиболее важная часть решения подобных задач - это выбор способа измерения расстояния между вершинами. Для этого используются различные метрики, которые отражают разные свойства графа. Наиболее простой способ определить расстояние - кратчайший путь - не всегда дает хорошие результаты, потому что этот метод не учитывает связи, которые длиннее, чем самая короткая, и их количество. Другая распространенная метрика - resistance distance, как и пропорциональная ей commute time distance, учитывает все возможные пути между вершинами. Однако, в работе [1] было показано, что при росте количества вершин в графе данные метрики сходятся к функциям, зависящим от степеней вершин и не отражающим глобальных свойств графа. Были предложены другие способы измерить расстояние между вершинами, большинство из которых представляет собой параметрические семейства и при предельных значениях параметров сходится либо к расстоянию кратчайшего пути, либо к resistance distance. В данной работе изучается поведение этих метрических семейств.

Целью работы является исследование зависимости метрик от параметров для трех типов случайных геометрических графов графов: ε-графов, графов ближайших соседей и полных графов с гауссовским распределением весов ребер.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Проанализировать существующую литературу по графовым метрикам
2. Разработать критерий сравнения метрик
3. Разработать модель, позволяющую генерировать графы и вычислять расстояния в терминах метрик
4. Исследовать поведение метрик в зависимости от параметров графа и параметров метрики
5. Вычислить оптимальные в смысле выбранного критерия качества значения параметров метрик для каждого из исследуемых типов графов

Глава 1

Постановка задачи

1.1 Основные определения

Пусть $G = (V, E)$ - неориентированный граф с множеством вершин V и множеством ребер E, n - число вершин. Матрицу смежности невзвешенного графа будем обозначать $A = (a_{ij})$, где $a_{ij} = 1$, если ребро $(v_i, v_j) \in E$ и $a_{ij} = 0$ в противном случае. Для взвешенных графов будем хранить в этой матрице веса ребер: $a_{ij} = w(v_i, v_j)$.

Расстояние между вершинами в графе задается матрицей расстояний $D = (d_{ij})$, которую получают из определенным образом заданных мер близости $H = (h_{ij})$ с помощью преобразования

$$D = (h\mathbf{1}^\intercal + \mathbf{1}h^\intercal - H - H^\intercal)/2,$$

где h - вектор-диагональ матрицы H.

В некоторых случаях вместо матрицы H используют матрицу H_0, состоящую из логарифмов элементов матрицы H.

НАЙТИ АНАЛОГИЧНЫЕ УТВЕРЖДЕНИЯ ДЛЯ НЕЛОГАРИФМИЧЕСКИХ ? Chebotarev - Graph bottleneck identity ?Buckley, Harari - Distance in graphs

Метрикой на множестве A называеся функция $d : A^2 \to \mathbb{R}$: такая, что для любых x,y,$z \in A$ выполнены следующие утверждения:

(1) $d(x,y) = 0$ тогда и только тогда, когда $x = y$

(2) $d(x,y) + d(x,z) - d(y,z) \geq 0$ (неравенство треугольника)

Из этого определения следует, что для любых x,$y \in A$

$d(x,y) = d(y,x)$ (симметричность)

$d(x,y) \geq 0$ (неотрицательность)

Рассмотрим другой класс функций, которые широко применяются в теории графов и сетей, исследовании марковских процессов и анализе статистических моделей.

Пусть A - непустое множество и $\Sigma \in \mathbb{R}$. Функция $\sigma : A^2 \to \mathbb{R}$ называется Σ-proximity на A, если для любых x,y,$z \in A$ выполняются следующие условия:

(1) $\Sigma_{t \in A}\sigma(x,t) = \Sigma$

(2) $\sigma(x,y) + \sigma(x,z) - \sigma(y,z) \leq \sigma(x,x)$, где при $z = y$ и $x \neq y$ неравенство строгое.

В работе [2] было доказано, что между метриками и Σ-proximities на множестве A существует взаимно однозначное соответствие.

Пусть G - мультиграф с набором вершин V. Функция $d : V * V \to \mathbb{R}$ называется cutpoint addictive (bottleneck addictive, graph-geodetic), если $d(i,j) + d(j,k) = d(i,k)$ выполнено тогда и только тогда, когда в графе G любой путь, соединяющий вершины i и k, проходит через вершину j.

Говорят, что матрица $S = (s_{ij}) \in \mathbb{R}^{n*n}$ задает транзитивную меру $s(i,j) = s_{ij}$ на вершинах $i,j \in V$ графа G, если ее элементы удовлетворяют транзитивному неравенству $s_{ij}s_{jk} \leq s_{ik}s_{jj}$.

Это неравенство является аналогом неравенства треугольника для мер близости.

Теорема Пусть $S = (s_{ij}) \in \mathbb{R}^{n*n}$ задает транзитивную меру на графе G и все недиагональные элементы этой матрицы положительны. Тогда матрица $D = (d_{ij})_{n*n}$, определенная как

$$D \ = \ (h\mathbf{1}^\intercal + \mathbf{1}h^\intercal - H - H^\intercal)/2,$$

где H получается поэлементным логарифмированием матрицы S, является матрицей расстояний на $V(G)$. Более того, это расстояние будет cutpoint addictive.

Доказательство этой теоремы можно найти в [1].

1.2 Задача

Пусть G - случайный геометрический граф. Требуется исследовать поведение параметрических семейств графовых метрик на этом графе и найти оптимальные параметры, при которых метрики дают наиболее полную информацию о структуре графа G.

Также требуется сравнить поведение логарифмических и нелогарифмических метрик.

В данной работе рассматривались три класса случайных геометрических графов: ε-графы, графы ближайших соседей, графы с гауссовским распределением весов ребер. Параметр (ε или количество соседей) выбирался таким образом, чтобы граф оказался связным с высокой вероятностью. Это делалось потому, что наибольший интерес для машинного обучения представляют именно связные графы.

Генерация графов происходила следующим образом: сначала случайным образом генерировались вершины (из равномерного распределения на d-мерном кубе и из смеси гауссовских распределений с различными параметрами смеси), затем по определенным правилам эти вершины соединялись ребрами. В случае ε-графов ребро (v_i, v_j) добавлялось в том случае, если евклидово расстояние между этими вершинами не превышало ε. В графах ближайших соседей каждая вершина соединялась ребром с k своими ближайшими соседями (TODO: у von Luxburg рассматривались 2 типа kNN-графов, symmetric и mutual). В гауссовских графах все вершины были соединены, а веса ребер определялись по формуле $w_{ij} = exp(-\frac{||v_i - v_j||^2}{\sigma^2})$, где параметр $\sigma > 0$

Для полученных графов вычислялись различные метрики, затем они сравнивались между собой. Для сравнения использовались коэффициент корреляции с евклидовым расстоянием и матричные нормы для матрицы $D_{euclid} - D_{metrics}$.

1.3 Исследуемые метрики

1. **Walk distance**

   Это параметрическое семейство строится с использованием меры близости $H = (I + tL)^{-1}$, где параметр $0 < t < \rho^{-1}$, ρ - спектральный радиус матрицы A. При предельных значениях параметра метрика сходится к shortest path distance и long walk distance.

2. **Logarithmic walk distance**

   Мера H_0 получается поэлементным логарифмированием матрицы H, определяющей Walk distance.

3. **e-walk distance**

4. **Forest distance**

   Данное семейство подробно описано в [3].

   Rooted tree - связный ациклический граф, одна вершина в котором отмечена как корень. Rooted forest - граф, все связные компоненты которого являются rooted trees.

   Рассмотрим взвешенный граф G. Обозначим $w(G)$ - произведение весов его ребер. Для графа без ребер $w(G) = 1$. Если S - набор графов, то $w(S) = \sum_{G \in S} w(G)$. В случае, когда S - пустое множество, $w(S) = 0$. Если множество S состоит из невзвешенных графов, то $w(S) = |S|$.

   Введем следующие обозначения: $F = F(G)$ - множество spanning rooted forests графа G; $F_{i,j} = F_{i,j}(G)$- множество таких spanning rooted forests, что вершина i принадлежит дереву с корнем j; $F_{i,j}^{(p)} = F_{i,j}^{(p)}(G)$ - подмножество таких spanning rooted forests множества $F_{i,j}$, которые содержат ровно p ребер.

   Пусть $f = w(F)$, $f_{i,j} = w(F_{i,j})$, $f_{i,j}^{(p)} = w(F_{i,j}^{(p)})$, где $i,j \in V(G)$ и $0 \leq p < n$.

   Рассмотрим матрицу $Q = (I + L)^{-1}$.

   Согласно matrix forest theorem, такая матрица существует для любого взвешенного мультиграфа и ее элементы равны $q_{i,j} = \frac{f_{i,j}}{f}$, где $i,j = 1,2...n$. Матрицу Q можно рассматривать как меру близости. Добавим зависимость от параметра:

   $H = (I + tL)^{-1}$, где параметр $t > 0$, L - лапласиан графа.

   При $t \to \inf$ данная метрика сходится к resistance distance. Доказательство этого факта можно найти в [11].(как и интерпретацию метрики)

5. **Logarithmic forest distance**

   H получена поэлементным логарифмированием матрицы близости для forest distance

6. **Communicability distance**

   [6]

   Communicability между вершинами p и q в графе G - это взвешенная сумма всех блужданий, которые начинаются в p и заканчиваются в q, при этом чем короче блуждание, тем больше его вес. Если A - матрица смежности графа, то Communicability между вершинами p и q - это соответствующий элемент матрицы e^A.

   Данное определение имеет простую физическую интерпретацию. Рассмотрим граф как систему из шариков массой m, соединенных пружинами с константой $m\omega^2$. Затем вся эта система погружается в жидкость с температурой T. Под воздействием температуры шарики начинают осциллировать.

   Гамильтониан системы имеет следующий вид:

   $H = \Sigma_i(\frac{p_i^2}{2m} + (K - k_i)\frac{m\omega^2 x_i^2}{2}) + \frac{m\omega^2}{2}\Sigma_{i,j:i<j}A_{ij}(x_i - x_j)^2$

   **********See Estrada p2**********88

   $H = e^{tA}$, параметр $t > 0$

7. **Logarithmic communicability distance**

   H получена поэлементным логарифмированием матрицы близости для communicability distance

8. **Free energy distance**

   Это семейство метрик, зависящее от параметра β, было рассмотрено в работе [7]. Расстояние вычисляется следующим образом:

   $P^{ref} = D^{-1}A$, $D = diag(Ae)$

   $W = P^{ref} * *e^{-\beta C}$, где C - матрица кратчайших расстояний между вершинами графа G

   $Z = (I - W)^{-1}$

   $Z^h = Z * D_h^{-1}$, $D_h = diag(Z)$

   $\Phi = -\frac{1}{\beta}\log Z^h$

   $D^{FE} = (\Phi + \Phi^T)/2$

   Данное расстояние стремится к расстоянию кратчайшего пути при $\beta \to \infty$ и к commute time при $\beta \to 0^+$

9. **Shortest path distance**

Кратчайшим путем между двумя вершинами графа называют такой путь между этими вершинами, что сумма весов ребер, из которых он состоит, минимальна.

Существует несколько способов вычисления кратчайшего пути, в данной работе используется алгоритм Флойда - Уоршелла [8].

10. **Resistance distance**

    Резисторное расстояние между двумя вершинами эквивалентно напряжению между соответствующими точками в электрической цепи, полученной из графа G заменой ребер на резисторы, сопротивление которых совпадает с весом ребер.

    $D = (L+J)^{-1}$, где L - лапласовская матрица, J - матрица, все элементы которой равны $\frac{1}{n}$, гдк n - число вершин

11. **Avrachenkov distance**

    Данное семейство мер близости было предложено в [9.]. Оно возникло при исследовании способов решения задачи классификации с частичным привлечением учителя (semi-supervised classification), которые основаны на использовании графов.

    $H = (1-a)(I - aD^{-\sigma}AD^{\sigma-1})^{-1}$, где $a = \frac{2}{2+\mu}$, μ - параметр регуляризации, который позволяет регулировать баланс между точностью классификации и гладкостью классифицирующей функции. Параметр σ позволяет использовать общую формулу для трех методов классификации с частичным привлечением учителя. При $\sigma = 1$ получаем метод, основанный на использовании стандартного лапласиана графа, $\sigma = 0.5$ - нормированного лапласиана, случай $\sigma = 0$ соответствует PageRank.

    D - матрица степеней вершин. В случае взвешенных графов вычисляется как сумма весов ребер, инцидентных данной вершине.

12. **Logarithmic Avrachenkov distance**

    Данная мера близости вычисляется с помощью поэлементного логарифмирования элементов матрицы H для метрики Авранченкова.

1.4 Формулы

При использовании дробей формулы могут получаться очень высокие:

$$\frac{1}{\sqrt{(2)} + \dfrac{1}{\sqrt{2} + \dfrac{1}{\sqrt{2} + \cdots}}}$$

В формулах можно использовать греческие буквы:

$$\alpha\beta\gamma\delta\epsilon\varepsilon\zeta\eta\theta\vartheta\iota\kappa\lambda\mu\nu\xi\pi\varpi\rho\varrho\sigma\varsigma\tau\upsilon\phi\varphi\chi\psi\omega\Gamma\Delta\Theta\Lambda\Xi\Pi\Sigma\Upsilon\Phi\Psi\Omega$$

1.4.1 Ненумерованные многострочные формулы

Вот так можно написать две формулы, не нумеруя их, чтобы знаки равно были строго друг под другом:

$$\begin{aligned} f_W &= \min\left(1, \max\left(0, \frac{W_{soil}/W_{max}}{W_{crit}}\right)\right), \\ f_T &= \min\left(1, \max\left(0, \frac{T_s/T_{melt}}{T_{crit}}\right)\right), \end{aligned}$$

Можно использовать разные математические алфавиты:

$$\mathcal{ABCDEFGHIJKLMNOPQRSTUVWXYZ}$$
$$\mathfrak{ABCDEFGHIJKLMNOPQRSTUVWXYZ}$$
$$\mathbb{ABCDEFGHIJKLMNOPQRSTUVWXYZ}$$

Посмотрим на систему уравнений на примере аттрактора Лоренца:

$$\begin{cases} \dot{x} = & \sigma(y - x) \\ \dot{y} = & x(r - z) - y \\ \dot{z} = & xy - bz \end{cases}$$

А для вёрстки матриц удобно использовать многоточия:

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

1.4.2 Нумерованные формулы

А вот так пишется нумерованая формула:

$$e = \lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n \tag{1.1}$$

Нумерованых формул может быть несколько:

$$\lim_{n\to\infty}\sum_{k=1}^{n}\frac{1}{k^2} = \frac{\pi^2}{6} \tag{1.2}$$

В последствии на формулы (1.1) и (1.2) можно ссылаться.

Глава 2

Исследование поведения метрик для различных типов графов

2.1 Генерации вершин графов

В данной работе вершины графа генерировались с помощью смеси гауссовских распределений. Основной случай:четыре двумерные гауссианы, центры которых расположены симметрично относительно начала координат, дисперсии и количество точек равны.

2.2 Генерации невзвешенных графов

В данной работе рассматривались три класса случайных геометрических невзвешенных графов графов:

1. **ε-графы**: вершины соединяются ребром в том случае, когда евклидово расстояние между ними не превышает заданного параметра ε.

2. **Симметричные графы ближайших соседей**: между двумя вершинами проводится ребро в том случае, если хотя бы одна из них попадает в множество k ближайших соседей другой; параметр k задан.

3. **Mutual графы ближайших соседей**: две вершины соединяются ребром, если обе они попадают в множество k ближайших соседей друг друга; параметр k задан.

Параметр графа (ε или k) выбирался таким образом, чтобы граф оказался связным с высокой вероятностью. Это делалось потому, что наибольший интерес для машинного обучения представляют именно связные графы.

ДОБАВИТЬ КАРТИНКИ С ПРИМЕРАМИ ГРАФОВ

а)

б)

в)

Рисунок 2.1: Примеры невзвешенных графов. Слева направо:

2.3 Генерации взвешенных графов

В данной работе взвешенные графы представлены гауссовскими графами. Это полные графы, в которых вес ребра между вершинами i и j определяется по формуле $w_{ij} = exp(-\frac{||v_i - v_j||^2}{\sigma^2})$, где параметр $\sigma > 0$ задан.

2.4 Сравнение метрик

Для каждого типа графов для различных значений параметра метрк вычислялись матрицы расстояний для каждой метрики, описанной в главе 1. Чтобы оценить «качество» метрик, они сравнивались с евклидовым расстоянием между вершинами графа. Для сравнения использовались коэффициент корреляции с евклидовым расстоянием и матричные нормы для матрицы $D_{euclid} - D_{metrics}$.

2.5 Одиночное изображение

Рисунок 2.2: TeX.

2.6 Пример вёрстки списков

Нумерованный список:

1. Первый пункт.
2. Второй пункт.
3. Третий пункт.

Маркированный список:

- Первый пункт.
- Второй пункт.
- Третий пункт.

Вложенные списки:

- Имеется маркированный список.
  1. В нём лежит нумерованный список,
  2. в котором
     - лежит ещё один маркированный список.

Глава 3

Результаты

3.1 Незвешенные графы

Результаты экспериментов представлены на графиках. Во всех случаях по оси x отложены значения параметра семейства. Для удобства все параметры были отнормированы на отрезок [0,1].

**Картинки **

Можем видеть, что логарифмические метрики позволяют более точно восстановить евклидово расстояние, чем метрики без логарифма.

Значения параметров, при которых метрики лучше всего приближают евклидово расстояние для каждого типа графов, приведены в таблице:

*****

про оптимальные степени

3.2 Взвешенные графы

На графиках представлены результаты сравнения метрик с евклидовым расстоянием для взвешенных Гауссовских графов.

******

Видно, что в данном случае расстояние кратчайшего пути позволяет восстановить евклидово расстояние с точностью до константы. Это связано с особенностями определения весов ребер.

Наилучшие значения параметров для остальных метрик приведены в таблице:

****

3.3 Таблица обыкновенная

Так размещается таблица:

Таблица 3.1: Название таблицы

Месяц	T_{min}, K	T_{max}, K	$(T_{max} - T_{min})$, K
Декабрь	253.575	257.778	4.203
Январь	262.431	263.214	0.783
Февраль	261.184	260.381	-0.803

Заключение

В данной работе было рассмотрено большое число параметрических семейств графовых метрик и исследовано их поведение в зависимости от параметра для четырех типов случайных геометрических метрик. Сравнение метрик осуществлялось посредством сравнения каждой из них с евклидовым расстоянием между вершинами графа.

На основании проведенных экспериментов можно сделать вывод, что логарифмическое преобразование позволяет значительно улучшить метрики.

Чиссленные исследования показали, какие параметры метрических семейств наиболее интересны для практических приложений, а также были вычислены степени, в которые можно возводить матрицу расстояний, чтобы получить лучшее приближение евклидового расстояния.

Исходный код для воспроизведения результатов, описанных в данной работе, доступен по адресу ********

Список литературы

1. Von Luxburg U., Radl A., Hein M. Hitting and commute times in large random neighborhood graphs //The Journal of Machine Learning Research. – 2014. – Т. 15. – №. 1. – С. 1751-1798.

1. P. Chebotarev. Studying new classes of graph metrics // F. Nielsen and F. Barbaresco (Eds.), Proceedings of the SEE Conference “Geometric Science of Information” (GSI 2013). Lecture Notes in Computer Science, LNCS 8085. Springer, Berlin, 2013. P. 207–214.

2. Chebotarev P. Y., Shamis E. V. On a Duality between Metrics and *Sigma*-Proximities //arXiv preprint math/0508183. – 2005.

2. Hitting and Commute Times in Large Random Neighborhood Graphs, U. von Luxburg, A. Radl, M. Hein. //Journal of Machine Learning Research 15 (2014) 1751-1798

3. A class of graph-geodetic distances generalizing the shortest-path and the resistance distances, P. Chebotarev. //Discrete Applied Mathematics 159 (2011) 295–302

4. Shortest path distance in random k-nearest neighbor graphs, M. Alamgir, U. von Luxburg (2012)

5. P.Chebotarev, E.Shamis. The forest metrics for graph vertices // Electronic Notes in Discrete Mathematics. 2002. V.11. P. 98–107.

6. E. Estrada. The communicability distance in graphs // Linear Algebra and its Applications. 2012. V. 436. P. 4317–4328.

7. I.Kivimäki, M. Shimbo, M. Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances // Physica A: Statistical Mechanics and its Applications. 2014. V. 393. P. 600–616.

8. R.W. Floyd. Algorithm 97: Shortest path. //Communication of the ACM 5(6):345, 1962.

9. K. Avranchenkov, A. Mishenin, P. Gonçalves, M. Sokol. Generalized Optimization Framework for Graph-based Semi-supervised Learning // In: Proceedings of SIAM Conference on Data Mining (SDM 2012)

10. Chebotarev Walk distances in graphs

11. Chebotarev. The forest metric for graph vertices.