Статистические модели для данных социальных сетей и использование байесовской методологии

И. Д. Калашников
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Financial University
kalashnikov-id@mail.ru

***Аннотация*. Байесовский подход в социальных науках был задуман как проявление, если таковые были необходимы, основных достижений в построении моделей, оценке и результатов, которые были достигнуты в Байесовской парадигме за последние несколько десятилетий. Эти достижения были неравномерными в различных областях, но, тем не менее, были широко распространены и идущими далеко вперед. Сегодня все отрасли общественных наук используют инструменты Байесовской статистики.**

Ключевые слова: Байесовский подход; стандарты; сетевые модели и моделирование; вероятностные оценки

I. Введение

В научных исследованиях люди, как правило, мотивированы и опираются на работу большого ряда других исследователей, чтобы произвести новые методы и выводы, которые в свою очередь формируют основу для дальнейшего изучения и научного дискурса. Междисциплинарные исследования могут быть особенно плодотворными в укреплении этой взаимосвязанности, несмотря на внутреннюю трудность тщательного плавания по незнакомым водам нескольких месторождений.

В частности, это связано с концептуальной простотой и интеллектуальной привлекательностью Байесовского подхода, но он имеет также много общего со способностью Байесовских методов к обработке ранее неразрешимых проблем из-за компьютерной революции, которая началась в 1990-х годах.

В современных экспериментах часто возникает ситуации, когда “классические” методы анализа погрешностей и доверительных интервалов дают неправильный результат. Обычно это связано с малой статистикой или близостью измеряемых величин к физически возможной границе.

- измерение массы нейтрино;
- изучение редкого процесса при наличии фона.

В подобных случаях байесовские методы оценки вероятностей приводят к более осмысленным результатам. Основные идеи были опубликованы в 1763 году в труде Томаса Байеса, поэтому эти методы такие же классические, как и “классические”. Однако широко байесовские подходы стали применяться только во второй половине двадцатого века.

Хотя конкретные темы и терминологии различаются, много общего можно найти в использовании новых современных вычислительных алгоритмов, сложном иерархическом моделировании и тщательном изучении неопределенности модели.

Изменение структуры и состава сети в последние два десятилетия представляло большой теоретико-методологический интерес, однако влияние эндогенных групповых изменений на динамику взаимодействия в контексте социальных сетей является удивительно недооцениваемой областью. Динамику сети можно рассматривать как процесс изменения структуры ребер сети, в наборе вершин, по которым определены ребра, или в обоих одновременно.

II. Статистические модели для данных социальных сетей

A. Сетевые данные и номенклатура

Следует сосредоточиться на сетях (социальных или иных), которые могут быть представлены с точки зрения дихотомических (т. е. неоцененных) связей между парами дискретных сущностей. Представляется набор из потенциально взаимодействующих объектов через набор вершин (V), с набором взаимодействующих пар (или упорядоченных пар, для направленного отношения), который представляет собой набор ребер (E). В комбинации эти два набора называются графиком, G = (V, E) (здесь, мы будем использовать термин «график» в общем виде для обозначения как направленных, так и ненаправленных структур, за исключением случаев, когда указано иное). Сети могут быть статичными, например, представляющими отношения в один момент времени или агрегированными в течение определенного периода времени, или динамическими, например, представляющими отношения, появляющиеся и исчезающие в непрерывном времени или в состоянии отношений через определенные дискретные интервалы времени.

Для многих целей, это полезно для представления графиков в терминах смежности матрицы для графика G порядка N = |V||, то матрица смежности Y {0, 1} N × N представляет собой матрицу индикаторных переменных, таких, как Yij = 1.

Следующие условности в социальной сети (но не график теоретический) литературе, мы будем называть N в качестве размера G.

Вышеуказанное естественным образом распространяется на случай динамических сетей в дискретное время. Рассмотрим временной ряд ..., Gt−1, Gt, Gt+1 ,..., где Gt = (Vt, Et), который представляет состояние системы интересов в момент времени t. Это соответствует, в свою очередь, матрице смежности серии ..., Y..t−1, Y..Y..t+1,..., с NT = |VT|. Размеры сети на время t и Y..T {0, 1} такие, что NtxNt Yijt = 1 примыкает к вершине GT в момент времени t.

B. Экспоненциальные модели случайных графиков

При моделировании социальных или других сетей часто полезно представлять их распределения через случайные графики в дискретной экспоненциальной форме семьи. Распределения графов, выраженные таким образом, называются экспоненциальными семейными случайными моделями графиков или ERGM. Холланд и Лейнхардт (1981), как правило, приписывают первое явное использование статистических экспоненциальных семей для отображения случайных графических моделей для социальных сетей. Сила этой структуры лежит в обширном теле теории инференциальных, вычислительных и стохастических процессов [заимствованных из общей теории дискретных экспоненциальных семейств], которые могут быть использованы для моделей, определенных в ее терминах. Мы начинаем со “статического” случая, в котором есть единичный случайный график, G, при поддержке G это удобно для модели G через его матрицы смежности Y, представляющий собой соответствующий (т. е., набор матрицы смежности, соответствующей всем элементам G). В форме ERGM, мы выражаем вероятностную массовую функцию из Y следующим образом:

$$\Pr(Y = y \mid S, \theta, X) = \frac{\exp\left(\theta^T S(y, X)\right)}{\sum_{y' \in y} \exp(\theta^T S(y', X))} Iy(y)$$

где $S : Y, x \rightarrow R^S$является вектором достаточных статистических данных, $\theta \epsilon R^S$ является вектором природных параметров, $X \epsilon\ X$ – коллекция ковариат, а функция Iy – индикатор (т.е., 1 если его аргумент в поддержку y, 0 – прототивоположно). Если |G| конечный, то PMF для любого G может очевидно написан с конечномерной с θ (например, сдача S - вектор индикатора переменных для элементов y); это не обязательно верно в более общем случае, хотя представление S, θ со счетным измерением все-таки существует. На практике, как правило, предполагается, что S имеет низкое измерение, или что, по крайней мере, вектор естественных параметров может быть отображен к низко-мерному вектору «изогнутых» параметров.

В то время как крайняя общность этой структуры сделала его привлекательным, выбор моделей и оценка параметров зачастую трудны из-за нормирующего коэффициента (к(θ,S, X) = Σy'∈ y $\exp\left(\theta^T S(y, X)\right)$ в знаменателе уравнения. Этот нормализующий фактор аналитически неразрешим и трудно вычислим, за исключением особых случаев, таких как Бернулли и диада многочлена семейства случайных графиков; первые приложения этого семейства сосредоточены на этих специальных случаях. Позже Фрэнк и Штраусс (1986) ввели более общую процедуру оценки на основе методов кумулянта, но это оказалось слишком нестабильным для практического использования.

C. Байесовский подход для ERGM параметров

С учетом вероятности в уравнении, Байесовский подход следует в обычном порядке с применением теоремы Байеса, т. е.

$$p(\theta|Y = y, S, X) = \frac{ERG(y|\theta, S, X)p(|\theta, S, X)}{\int_R^0, ERG(y|\theta, S, X)p(|\theta, S, X)d\theta'}$$

$$\infty\ ERG(y|\theta, S, X)p(y|\theta, S, X),$$

где $p(\theta|Y = y, S, X)$ является апостериорной плотности θ с учетом наблюдаемого состояния Y, статистика вектора S, и ковариаты X, p (θ|S, X) – соответствующая плотность до тета на R^S, и $ERG(y|\theta, S, X)$ представляет вероятность ERGM для PR(Y = y|θ, S, X) .

III. Байесовский подход для параметров DNR

Поскольку семейство динамической сетевой логистической регрессии (DNR) DNR сводится к структуре логистической регрессии, байесовский подход номинально прост. Однако выбор прежней структуры для семей DNR до сих пор не изучен. Правомерно или иным образом, исследователи, как правило, стремятся использовать предыдущую спецификацию по умолчанию, если они не имеют сильного обоснования для утверждения конкретного предшествующего. Есть обширная литература по неинформативным, дефолтным и ссылочным предшествующим распределениям в Байесовском статистическом поле. Один все более широко используемый подход к оценке приоритетов по умолчанию (особенно в литературе машинного обучения) заключается в использовании следующих методов:

прогнозирующая оценка, т. е. изучение степени надежности данной предшествующей структуры приводит к точным прогнозам по тестовым данным для заданного тела обучающих данных. Хотя, возможно, приоритеты, найденные, чтобы дать хорошую предсказательную производительность на прошлых данных могут быть привлекательными на прагматических основаниях; в свою очередь, такие приоритеты могут быть также оправданы более предметно, как представляющие дистрибутивы совместимы с прошлых наблюдений на аналогичных данных, и, следовательно, менее правдоподобно качество отправной точки. Аналогичным образом, следует подозревать приоритетов, которые последовательно приводят к плохой прогностической производительности тестовых данных, независимо от принципов, используемых для их создания. Таким образом, баланс касается прогнозной оценки различных кандидатов в президенты в контексте моделей DNR.

В то время как MCMC и MAP возможны для подхода в семьях DNR, наш фокус здесь будет сконцентрирован на задней панели моделирования по схеме. В дополнение к этому она дает нам более полное представление о заднем распределении, заднее моделирование особенно хорошо адаптированы к прогностической модели проверки адекватности. В частности, моделирование будущих наблюдений условно на оценке пункта (например, заднем или среднем режиме), может значительно недооценивать неопределенность, связанную с задним распределением, и расширение может не выявить преимуществ, которые должны быть получены, например, предыдущие спецификации, которые преобладают в экстремальных значениях параметров без существенного изменения центральной тенденции заднего распределения.

Учитывая вышесказанное, существует много разумных вариантов предыдущих спецификаций для семей DNR, которые могут быть применимы в том или ином контексте. Учитывая, что наше внимание сосредоточено на простой оценке, приоритете значения по умолчанию, мы сфокусируем наше внимание на 4 предыдущих спецификациях, предложенных как по умолчанию для логистической регрессии в байесовской статистической литературе. Наши основные вопросы следующие.

Во-первых, каковы последствия использования этих эталонных приоритетов в сравнении с оценкой максимальной вероятности для семей DNR в типичные настройках социальных сетей? Во-вторых, в какой степени делают различные разумные приоритеты дефолта, что приводит к различиям в оценке точки или задней неопределенности в таких условиях? Наконец, какие различия (если таковые имеются) делает выбор того или иного значения по умолчанию до создания, чтобы предсказать специфическое чувство прогнозирования свойств развивающейся сети? Если, в обычных настройках, дедуктивные прогнозирования довольно чувствительны к выбору до, то выбор на основе вычислительных или других факторов может быть разумной практикой. Если, напротив, мы найдем существенные различия в косвенных и/или прогнозных показателей среди настоятелей по умолчанию, а затем эти варианты должны быть тщательно изучены. Баланс данной главы предназначен для первого шага к оценке этих вопросов.

Хотя моделирование и подход для ERGM в целом является узкоспециальным искусством, логистическая форма семей DNR облегчает оценку параметров (если не сетевое моделирование) с использованием более стандартизированных инструментов и методов.

IV. Байесовская оценка DNR с вершиной

Как отмечалось выше, в данном случае делимости ребра и вершины процессы в DNR контекст допускает, что обе должны рассматриваться как независимые проблемы логистической регрессии (пока связанные параметры априори независимы). А это означает, нет повода, что до структуры должны использоваться для обоих; для простоты и практичности, мы рассмотрим случай, когда мы предполагаем, что обе кромки и параметры вершин имеют одинаковые априорные распределения.

Как отмечалось выше, здесь рассматривается ряд типичных приоритетов логистической регрессии (рекомендованных в литературе) для использования в качестве приоритетов в DNR с динамикой вершин и без нее. Базовой точкой сравнения для всех байесовских результатов будет оценка ML (и его распределение по выборке), отражающая доминирующую практику в литературе ERGM. В дополнение к этому базовому уровню мы рассмотрим пять предыдущих спецификаций в трех общих классах:

1. Неправильная равномерная предварительная оценка (т. е. полностью байесовский аналог максимальной вероятности).
2-3. Независимые нормально распределенные приоритеты (один центрированный на 0, и одно смещение для завышения предыдущей плотности).
4-5 Слабо информативные семейства правильных приоритетов. Эмпирические эксперименты по воздействию этих приоритетов на анализ и интерпретацию можно найти в следующем разделе.

V. Эмпирические примеры и анализ моделирования

Учитывая шесть описанных выше спецификаций (включая MLE), мы стремимся оценить практические последствия этих приоритетов для байесовского подхода в типичных условиях. С этой целью мы рассматриваем сравнительный анализ подходов в рамках предложенных нами приоритетов по двум эмпирическим случаям. Первая – это динамичная сеть цитат среди блогеров во время президентских выборов 2004 года в США («Блог»), а вторая – динамичная сеть лицом к лицу коммуникационных связей среди виндсерферов на пляже Южной Калифорнии («пляж»). Обе сети типичны для наборов данных социальных сетей, в каждой из которых задействованы многочисленные, сложные механизмы взаимодействия, а также гетерогенность на уровне субъектов. Эти сети также изучены в литературе, что делает их полезными справочными примерами для нашего настоящего анализа.

Данные блога

Первый набор данных, рассмотренных в этой главе, представляет собой динамический межгрупповой блог цитирования сети. Это динамическая сеть состоит из взаимодействия между всеми блогов аккредитованным Национальным Комитетом Демократической партии (НКДП) или Республиканского Национального комитета (РНК) по соответствующим конвенциям 2004 года. Множество вершин состоит из субъектов, представляющих 14 аккредитованных блогов, а также блогов уполномоченных обеих групп, и является статическим в течение периода наблюдения. Краевой набор состоит из направленных гиперссылок на первых страницах вышеупомянутых блогов, наблюдаемых в течение 121 дня. Данные были получены путем

автоматического запроса главной страницы каждого блога с шестичасовыми интервалами, начиная с полуночи по тихоокеанскому времени. Период исследования начинается 22.07.04 (незадолго до принятия Конвенции DNR) и заканчивается 19.11.2004 (вскоре после президентских выборов), что в общей сложности приводит к 484 моментам времени. Существует ребро из блога в блог J в момент времени t, появится ссылка на блог J на главной странице в момент времени t. В частности, эти данные представлены как массив смежности.

VI. Современный Байесовский факторный анализ

Происхождение факторного анализа можно проследить в работе Спирмена (1904), посвященной общей разведке. В свое время психологи пытались определить интеллект единой, всеохватывающей ненаблюдаемых сущностей, G-фактора. Спирмен исследовал влияние G-фактора на результаты тестирования испытуемых на нескольких доменах: поле, свету, весе, классике, французском языке, английском языке и математике. В конце концов, G-фактор будет обеспечивать механизм и обнаружит общие корреляции между такими несовершенными измерениями. Точнее, однофакторная модель Спирмена (1904), основанная на p – тестовых доменов (измерений) и N – испытуемых (физических лиц) может быть записана как

$$y_{ij} = \mu_j + \beta_j g_i + \varepsilon_{ij},$$

для i = 1,..., n и j = 1,..., p, где y_{ij} – это результат испытуемого i на тестовом домене j, μ_j – это проверка домена i, g_i – это величина фактора интеллекта для субъекта, β_j – это ожидание проверки влияния домена j на фактор интеллекта g, ε_{ij}, – это случайные ошибки субъекта i и проверка домена j.

Дж. Спирмен тратит часть своего 90-страничного документа, защищая его однофакторную модель общего интеллекта, что, пожалуй, стало его главным, основополагающим вкладом в область психометрии, а также статистическое моделирование в это время. Тем не менее, Спирмен изобрел факторный анализ, но его почти исключительную озабоченность понятием общего фактора помешала ему реализовать свой потенциал в полном объеме.

Распространение на многочисленные факторы, а также формальные статистические рамки произошло спустя много десятилетий. Многофакторный анализ впервые был введен Thurstone (1935, 1947) и Lawley (1940, 1953), наряду с оценкой с помощью центроидного метода и максимальной вероятностью, соответственно.

A. Смесь факторных анализаторов

Смесь факторных анализаторов (СФА) представляет собой нелинейное, более гибкое расширение линейного факторного анализа раздела. Базовая структура модели СФА дается выражением

$$\left(y\middle|\beta, f, \sum y\right) \sim \sum_{j=1}^{m} \pi_j N\left(y; \beta_j f; \sum y\right),$$

где m – количество анализаторов. Условно на j возникает стандартная нормальная линейная факторная модель. Был изобретен алгоритм, который содержит смесь факторных модели аналитиков с помощью вариационного приближения к полной Байесовской интеграции над параметрами модели.

VII. Заключение

Байесовский анализ обеих задач в обобщенной базе затрудняется центральной ролью несовместимого фактора нормализации ЭРГМ, который входит в вероятность (а иногда и предыдущую) таким образом, что делает традиционную основу MCMC схемы отбора проб медленной и / или непрактичной. Однако для некоторых семейств моделей эта проблема не применяется (из-за наличия трактуемого нормализующего фактора); такие семейства особенно полезны в случае динамического сетевого моделирования, когда кондиционирование в прошлом может в некоторых случаях позволить моделировать ребра как условно независимые в настоящем. В таких случаях временная форма ЭРГМ сводится к простому произведению неоднородных подобий графика Бернулли, получившем название “динамическая сетевая регрессия” из-за сходства получаемой модели с логистической регрессией по данным временных рядов. Наши эксперименты с двумя разными наборами данных свидетельствуют о том, что это разумный подход: альтернативные критерии дефолта приводят к очень схожим выводам, причем все они сходны с выводами, вытекающими из максимальной подобной оценки. Одной из причин такого согласия является то, что даже довольно скромные динамические сетевые наборы данных обеспечивают достаточное количество степеней свободы данных, чтобы – для семейств DNR – поместить заднюю часть в асимптотический Гауссовский режим.

Учитывая это, а также преимущества байесовского подхода к таким проблемам, как прогнозирование, представляется маловероятным не рекомендовать его в качестве стандартного метода анализа динамики сети с семействами TERGM DNR.

Список литературы

[1] Айзерман М.А., Браверман Э.М., Розоноэр Л.И. «Метод потенциальных функций в теории обучения машин». Наука, 1970.

[2] Ветров Д.П., Кропотов Д.А. «Алгоритмы выбора моделей и синтеза коллективных решений в задачах классификации, основанные на принципе устойчивости» УРСС, 2006.

[3] Звягин Л.С. Процесс обработки информации при реализации концепции "мягких" измерений // Международная конференция по мягким вычислениям и измерениям. 2017. Т. 1. С. 104-109.

[4] Звягин Л.С. Применение системно-аналитических методов в области экспертного прогнозирования // Экономика и управление: проблемы, решения. 2017. Т. 3. № 6. С. 145-148.

[5] Звягин Л.С. Проблемы внедрения системного анализа в целевом управлении // Всероссийская научная конференция по проблемам управления в технических системах. 2017. № 1. С. 305-308.

[6] D. MacKay Information Theory, Inference, and Learning Algorithms Cambridge University Press, 2003.

[7] V.N. Vapnik «The Nature of Statistical Learning Theory Springer», 1995.