Копулы в задачах оценки интенсивности рискованного поведения индивида по данным о последних эпизодах поведения [1]

В.Ф. Столярова аспирант, м.н.с лаборатории теоретических и междисциплинарных проблем информатики, СПИИРАН
valerie.stoliarova@gmail.com

Аннотация

Ряд задач эпидемиологии и охраны общественного здоровья связан с необходимостью оценки интенсивности рискованного поведения индивида по сверхкоротким данным о его поведении: по данным о последних эпизодах. Математические модели поведения индивида позволяют строить такие оценки. Практическое применение таких моделей основано на применении аппарата байесовских сетей доверия. В работе предложен другой класс моделей для построения требуемой оценки: класс непараметрических непрерывных байесовских сетей доверия. Вершины этой вероятностной графической модели представляют собой непрерывные переменные, связанные между собой копулами.

Введение.

Ряд задач эпидемиологии и охраны общественного здоровья [10] тесно связан с оценкой вреда, который может быть причинен индивидом обществу, самому себе и/или другому индивиду. В таком случае с риском связывают эпизоды определенного поведения индивида, а численной характеристикой такого риска выступает интенсивность поведения. Однако прямая оценка интенсивности социально-значимого поведения не всегда доступна в силу экономических причин или же свойств памяти респондентов. Для построения оценки интенсивности поведения по данным о последних эпизодах были предложены пуассоновская и гамма-пуассоновская математические модели поведения.

[1] Статья содержит материалы исследований, частично поддержанных грантами РФФИ 14-01-00580, 15-01-09001-а.

Практическое применение предложенных моделей основано на построении байесовской сети доверия (БСД) [8]. Байесовские сети доверия [9] представляют собой гибкий аппарат, которые позволяет как строить оценки интенсивности по имеющимся данным, так и обучать числовые и графические параметры модели по новым данным. Однако полная спецификация такой модели требует задания априорных значений условных вероятностей, которые сложно установить экспертными методами, а статистических данных может быть недостаточно для оценки большого числа параметров. Кроме того, дискретизация непрерывных по своей природе длин интервалов приводит к потере информации.

Непараметрические непрерывные байесовские сети доверия (ННБСД) [2] позволяют избежать дискретизации переменных. Вместо задания тензоров условных вероятностей, достаточно указать значение коэффициента (условной) ранговой корреляции между переменными модели, связанными отношением ребенок–родитель. Эта вероятностная графическая модель во многом опирается на аппарат копул [4].

В работе построена и программно реализована модельная ННБСД для оценки интенсивности поведения по данным о последних эпизодах.

Теоретические основы.

Копулы. Параметризации многомерных вероятностных распределений при помощи копул.

Пусть имеются n случайных величин $X_1, X_2, \ldots, X_n$, имеющих распределения вероятности $F_k(x_k) = P[X_k \le x_k]$, $k = 1 \ldots n$. Пусть $H(x_1, \ldots, x_n) = P[X_1 \le x_1, \ldots, X_n \le x_n]$ совместная функция распределения этих случайных величин. В этом случае функции F_k называются *маргиналами* совместной функции распределения H. Таким образом, с каждой точкой $(x_1, x_2, \ldots, x_n)$ связаны $n+1$ чисел, лежащих на отрезке $I = [0, 1]$: $F_k(x), k = 1 \ldots n$ и $H(x_1, \ldots, x_n)$. Копулы представляют собой функции, которые позволяют разделить совместную функцию распределения $H(x_1, \ldots, x_n)$ на маргинальную составляющую и составляющую собственно зависимости между переменными.

Понятие копулы может вводится с функциональной и с вероятностной точек зрения.

Определение 1 *(n-мерная копула). n-мерной копулой называется* *функция $C : I^n \to I$, такая что*

1. *для каждого*

$$\mathbf{u} \in I^n, \text{ такого что } \exists k : u_k = 0, C(u) = 0;$$

2. *для каждого*

$$\mathbf{u} \in I^n, \text{ такого что } \exists k : u_k \neq 1, \forall j \neq k u_j = 1, C(u) = u_k;$$

3. *для каждых $\mathbf{a} = (a_1, \ldots, a_n), \mathbf{b} = (b_1, \ldots, b_n) \in I^n : \mathbf{a} \leq \mathbf{b}$:*

$$\Delta_{a_n}^{b_n} \ldots \Delta_{a_k}^{b_k} \ldots \Delta_{a_1}^{b_1} C(\mathbf{t}) \geq 0,$$

где символом $\Delta_{a_k}^{b_k} C(\mathbf{t})$ обозначена разность значений функции C в точках a_k и b_k по координате с индексом k:

$$\Delta_{a_k}^{b_k} C(\mathbf{t}) =$$
$$C(t_1, \ldots, t_{k-1}, b_k, t_{k+1}, \ldots, t_n) - C(t_1, \ldots, t_{k-1}, a_k, t_{k+1}, \ldots, t_n)$$

С вероятностной точки зрения:

Определение 2 *n–копула представляет собой ограниченное на единичный куб I^n совместное распределение вероятности n случайных величин с равномерным распределением на отрезке I.*

Действительно, n-мерная функция распределения удовлетворяет свойствам из 1 и наоборот [6]. Напомним, что для любой непрерывной случайной величины X с распределением вероятности $F(x)$, случайная величина $U = F(X)$ будет иметь равномерное распределение.

Роль, которую копулы играют в описании взаимосвязи переменных, устанавливает теорема Склара (как и определение копулы, теорема Склара может быть сформулирована в функциональной и в вероятностной интерпретации; для наглядности приведем вероятностную формулировку теоремы).

Теорема 3 (Теорема Склара) *Пусть имеются n случайных величин $X_1, X_2, \ldots, X_n$, имеющих распределения вероятности F_k, $k = 1 \ldots n$ и совместную функцию распределения H. Тогда существует n–копула C такая что для всех $\mathbf{x} \in \bar{\mathbf{R}}^n$*

$$H(x_1, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)). \tag{1}$$

Если функции F_k все непрерывны, то C единственна; иначе, C однозначно определена на $\mathrm{Ran}F_1 \times \mathrm{Ran}F_2 \times \ldots \times \mathrm{Ran}F_n$*. Обратно: если C есть n–копула и $F_k, k = 1 \ldots n$ есть функции распределения, тогда определенная выше (1) функция $H(x_1, \ldots, x_n)$ является совместной функцией распределения.*

Таким образом, копулы позволяют разделить взаимосвязь двух переменных на маргинальную составляющую и составляющую зависимости между переменными. Такое представление удобно при моделировании, так как позволяет разделить задачи оценивания параметров. К примеру, если структура взаимосвязи (вид копулы C) переменных модели известна заранее, скажем, в ходе предыдущих фаз экперимента, то по данным требуется оценить лишь параметры маргиналов. Меньшее количество параметров требует меньших вычислений и меньшего объёма выборки.

Копулы действительно содержат информацию о зависимости между переменными, так как они инвариантны относительно монотонных преобразований (или изменяются предсказуемым образом) [4]. Вторым основополагающим свойством копул является *липшицевость* копул [5].

Однако построение могомерных копул представляет собой достаточно сложную задачу; в частности, если подставить в выражение для копулы многомерные функции распределения, то не всегда возможно получить снова копулу. Это свойство чаще всего выполняется для важного класса *архимедовых копул* [4].

Практически при моделировании многомерных распределений вероятности копулы соединяются попарно, при этом структура зависимостей между переменными описывается специальным объектом: *лозой (vine)* [2].

Определение 4 *Лоза над n элементами представляет собой множество деревьев $N = (T_1, \ldots, T_{n-1})$, причем ребра каждого дерева j являются вершинами дерева $j + 1$ и каждое дерево имеет наибольшее число ребер.*

Копулы и модели лозы играют важную роль при построении непрерывных непараметрических байесовских сетей доверия [2]. Этот класс моделей имеет в основе направленный ациклический граф, с вершинами, как и в случае лозы, связаны непрерывные случайные величины. Ребра представляют собой причинно-следственные связи между переменными, численным параметром выступает коэффициент (условной) ранговой корреляции.

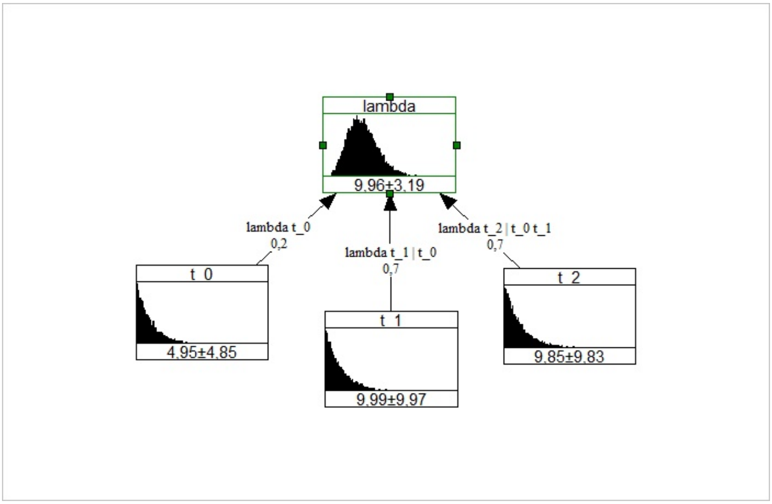

Рис. 1: НПНБСД для задачи оценки интенсивности по данным о трех последних эпизодах поведения

НПНБСД в задаче оценки интенсивности поведения по данным о последних эпизодах

Пусть имеется последовательность интервалов между эпизодами поведения $\tau_1, \ldots, \tau_k$. Пусть, согласно имеющимся моделям поведения индивида, эти переменные независимы и имеют экспоненциальное распределение вероятности с $\lambda = 10$. Пусть длина интервала между моментом интервью и последним эпизодом поведения не зависит от остальных интервалов и имеет экспоненциальное распределение вероятности с параметром $\lambda = 5$. Все интервалы связаны с параметром интенсивности поведения индивида, имеющей гамма-распределение с параметрами $\alpha = 1, \beta = 10$. Тогда структура непараметрической непрерывной байесовской сети доверия имеет вид:

Для спецификации модели необходимо задать значения (условных) ранговых корреляций переменных, соединенных ребром (конкретные значения отображены на рисунке [1]). Напомним, что если известен тип копулы (наиболее распространена нормальная копула [2, 3]) и значение коэффициента ранговой корреляции, то параметр копулы опре-

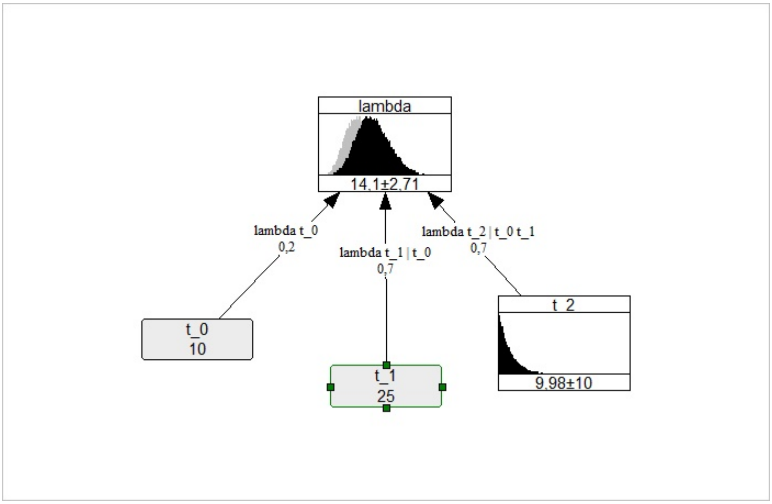

Рис. 2: Пример расчета интенсивности по данным о дввух последних эпизодах поведения

деляется однозначно. В частности, для нормальной копулы τ Кендалла вычисляется по формуле [3]:

$$\tau(\theta) = 2/\pi \arcsin(\theta). \tag{2}$$

Таким образом, при задании конкретных значений переменных модели,происходит распространение свидетельства по ребрам. Пусть, к примеру, наблюдаются следующие значения длин интевалов: $\tau_0 = 10$, $\tau_1 = 25$. Тогда, согласно имеющейся модели, плотность вероятности переменной λ изменится. Значения коэффициентов ранговой корреляции и На рисунке [2] серым контуром показана гистограмма частот до введения данных, а черным — после.

Для моделирования непараметрической непрерывной байесовкой сети доверия было использовано программное обеспечение UniNet [11].

Список литературы

[1] Barlow R. E., Proschan F. Mathematical Theory of Reliability. Classics of Aplied Mathematics, vol. 17. SIAM, 1996. 258 p.

[2] Kurowicka D., Joe H. (eds.) Dependence modeling: vine copula handbook. World Scientific Publishing Co, 2011. 370 p.

[3] Meyer C. The bivariate normal copula // Communications in Statistics-Theory and Methods, 2013. Т. 42, № 13. Стр. 2402-2422.

[4] Nelsen R. B. An introduction to Copulas, second edition. Springer series in Statistics, 2006. 272 p.

[5] Благовещенский Ю. Н. Основные элементы теории копул // Прикладная эконометрика, 2(26). 2012. Стр.113–130.

[6] Гнеденко Б. В. Курс теории вероятностей. 8-е издание. М.: Едиториал УРСС, 2005. 448 с.

[7] Степанов Д. В., Мусина В. Ф., Суворова А. В., Тулупьев А. Л., Сироткин А. В., Тулупьева Т. В. Функция правдоподобия с гетерогенными аргументами в идентификации пуассоновской модели рискованного поведения в случае информационного дефицита // Тр. СПИИРАН, 2012. № 23. Стр. 157–184. URL: http://mi.mathnet.ru/trspy557

[8] Суворова А. В., Тулупьева Т. В. , Тулупьев А. Л., Сироткин А. В., Пащенко А. Е. Вероятностные графические модели социально-значимого поведения индивида, учитывающие неполноту информации // Труды СПИИРАН, 2012. № 3, стр. 101–112.

[9] Тулупьев А. Л., Николенко С. И., Сироткин А. В. Байесовские сети: логико-вероятностный подход. СПб. : Наука, 2006.

[10] Тулупьева Т. В., Пащенко А. Е., Тулупьев А. Л., Красносельских Т. В., Казакова О. С., Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей, 2008, 140 с.

[11] URL: http://www.lighttwist.net/wp/uninet (доступ 12.06.2016)