Способы восстановления пропущенных значений в выборках из многомерных распределений с использованием марковских цепей

Бушмакина Анна Владимировна
bbushmakina.a@gmail.com

МГУ имени М.В. Ломоносова

Научный руководитель: профессор Яровая Елена Борисовна

26 мая 2023 г.

Мотивация и цели работы

Мотивация:

- Методы заполнения, основанные на случайных лесах (RF), не требуют настройки параметров модели и эффективны для задачи заполнения пропущенных значений.
- Реализация метода восстановления данных с использованием марковских цепей (MICE, S.Van Buuren, 2011) на случайных лесах может повысить точность заполнения по сравнению с другими методами.
- Однако до сих пор неясно, как методы заполнения, основанные на RF работают с ненормально распределенными данными и нелинейными зависимостями между переменными.

Цели:

1. На основании теории марковских цепей реализовать метод MICE на RF.
2. Исследовать MICE и его эффективность при восстановлении данных с переменной, распределение которой отличается от нормального, а также при наличии в данных нелинейных отношений между переменными.
3. Сравнить метод MICE с другими методами для выявления наиболее эффективного с точки зрения заданных метрик.
4. Разработать методику применения MICE для заполнения пропусков в реальных данных.

Необходимые сведения из теории марковских цепей

Теорема (Эргодическая теорема. (М.Кельберт, Ю.Сухов, 2017))

Если конечная цепь Маркова с переходной матрицей $\boldsymbol{P} = (p_{i,j})_{i,j \in S}$ *является неразложимой и апериодичной, то при любом начальном распределении* $\lambda = (\lambda_1, \dots, \lambda_N)$ *вероятности* $P(X_n = i)$ *сходятся к некоторым* π_i*,* $i = 1, \dots, N$ *при* $n \to \infty$*:* $\boldsymbol{P}^n \to \ \boldsymbol{\Pi}, n \to \infty$*.*

При этом вектор $\pi = (\pi_1, \dots, \pi_N)$ *является единственным решением системы уравнений* $\pi \boldsymbol{P} = \pi$ *и называется эргодическим (стационарным, инвариантным) распределением.*

Таким образом, с течением времени для неприводимой апериодичной цепи Маркова получаем распределение, не зависящее от первоначального:

$$\lim_{n\to\infty} P(X_n = j) = \lim_{n\to\infty} \sum_i \lambda_i p_{ij}^{(n)} = \sum_i \lambda_i \lim_{n\to\infty} p_{ij}^{(n)} = \pi_j. \tag{1}$$

Гиббсовское семплирование

Идея

Иногда сложно находить совместное распределение двух признаков $p(x, y)$, и проще перейти к нахождению условных вероятностей $p(x|y)$, $p(y|x)$.

Семплирование

- некоторое возможное значение x_0 для x
- используем x_0 для генерации y_0 : $y_0 \sim p(y|x = x_0)$
- используем y_0 для генерации x_1 : $x_1 \sim p(x|y = y_0)$
- и так далее: $x_i \sim p(x|y = y_{i-1}), y_i \sim p(y|x = x_i)$

Таким образом, последовательность сгенерированных значений $(x_0, y_0), (x_1, y_1), \ldots$ образует *неприводимую и апериодичную* цепь Маркова (Bendimerad A. et al., 2020), *эргодическое* распределение которой является искомым.

Множественное заполнение методом цепных уравнений

В основе MICE лежит алгоритм гиббсовского семплирования:

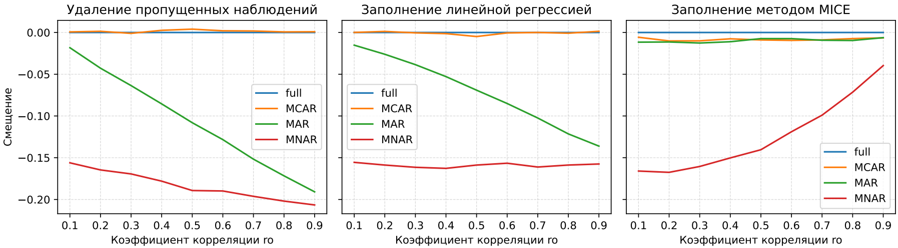
Шаг 0. Для каждого Y_j:

1. Указать модель заполнения $P(Y_j^{mis}|Y_j^{obs}, Y_{-j})$;
2. Выбрать первоначальное заполнение Y_j^0 случайным образом из Y_j^{obs};

Итерации $t = 1, \ldots, N$:

1. Определяем $Y_{-1}^t = (Y_2^{t-1}, \ldots, Y_p^{t-1})$,
2. $\theta_1^t \sim P(\theta_1^t|Y_1^{obs}, Y_{-1}^t)$,
3. $Y_1^t \sim P(Y_1^{mis}|Y_1^{obs}, Y_{-1}^t, \theta_1^t)$,
4. $\ldots$
5. $Y_{-j}^t = (Y_1^t, \ldots, Y_{j-1}^t, Y_{j+1}^{t-1}, \ldots, Y_p^{t-1})$,
6. $\theta_j^t \sim P(\theta_j^t|Y_j^{obs}, Y_{-j}^t)$,
7. $Y_j^t \sim P(Y_j^{mis}|Y_j^{obs}, Y_{-1}^t, \theta_j^t)$,
8. $\ldots$
9. Переход на следующую итерацию $t + 1$

Сравнение распространенных методов заполнения на данных, имеющих нормальное распределение



- Данные, потерянные неслучайным образом (MAR, MNAR) плохо восстанавливаются распространенными методами
- При удалении и заполнении линейной регрессией с ростом корреляции в данных увеличивается смещение
- Смещение данных с потерей MAR, заполненных методом MICE близко к 0. Смещение данных с потерей MNAR уменьшается с ростом корреляции

Заполнение методом MICE на данных, распределение которых отличается от нормального

4 сценария взаимосвязи между переменными, где $\varepsilon \sim N(0,1)$:

1. линейная регрессия с квадратичным членом:
   $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Пример: $Y = 2 + 2X + X^2 + \varepsilon$.
2. логистическая регрессия с квадратичным членом:
   $Y \sim Binomial(1, \alpha)$, $logit(\alpha) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
   Пример: $logit(\alpha) = -1.2 + 0.1X + 0.05X^2 + \varepsilon$.
3. линейная регрессия с членом XZ:
   $Z \sim Normal(4, 2)$, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.
   Пример: $Y = 2 + X + XZ + Z + \varepsilon$.
4. логистическая регрессия с членом XZ:
   $Z \sim Normal(4, 2)$, $Y \sim Binomial(1, \alpha)$,
   $logit(\alpha) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.
   Пример: $logit(\alpha) = -2 + 0.5X - 0.0625XZ + 0.25Z + \varepsilon$.

Заполнение методом MICE на данных, распределение которых отличается от нормального

Распределения X:

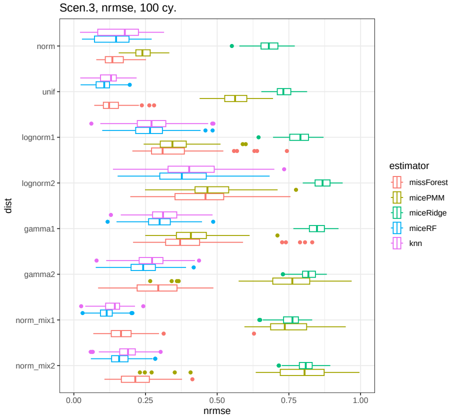
1. Normal(4, 1)
2. Uniform(0, 8)
3. Lognormal(0, 0.25)
4. Lognormal(0, 0.625)
5. Gamma(1, 1)
6. Gamma(2, 0.5)
7. $[N(1,1), N(6,3)]$
8. $[N(1,1), N(6,10)]$

Способы сравнения точности заполнения данных

1. Нормированная среднеквадратичная ошибка (S.Oba et al., 2003):
   $\sqrt{\frac{mean((X_{true}-X_{imp})^2)}{var(X_{true})}}$, где X_{true} — исходные данные, X_{imp} — заполненные.
2. Относительное смещение среднего значения заполненной переменной
   $\frac{mean(V_{imp})}{mean(V_{true})} - 1$, где V – одна из переменных $\{X, X^2, XZ\}$.
3. Относительное смещение оценки коэффициента $(\widehat{\beta}_l - \beta_l)/\beta_l$, $l = \{1, 2\}$.

Заполнение методом MICE на данных, распределение которых отличается от нормального. Результаты

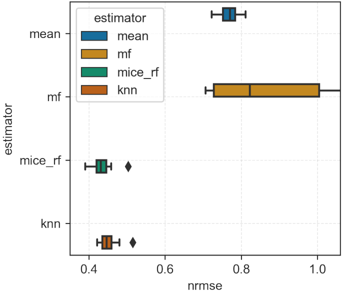
- Моделирование на сгенерированных данных показало, что miceRF имеет наилучшие результаты с точки зрения нормированной среднеквадратичной ошибки, относительного смещения среднего значения заполненной переменной и относительного смещения оценки коэффициента регрессии.
- Для логистических регрессионных зависимостей miceRidge и knn также являются эффективными методами заполнения



Применение метода MICE к реальным данным (Г.А.Муромцева et al., 2014)

Схема работы с пропущенными данными:

1. Исследуем, в каких переменных присутствуют пропущенные значения.
2. Удаляем наблюдения (строки) с пропущенными значениями.
3. В полном датасете симулируем потерю данных (метод *ampute* (R.M.Schouten, et al., (2018)) по переменным из шага (1).
4. Применяем методы заполнения и сравниваем результаты заполнения по выбранной метрике.
5. Заполняем исходный набор данных методом, показавшим наилучшие результаты.



Научная новизна и практическая ценность

Научная новизна

- Исследование эффективности метода MICE на различных моделях (случайные леса, ридж регрессия)
- Сравнение методов в наборах данных с переменной, распределение которой отличается от нормального, а также при наличии в данных нелинейных отношений между переменными
- Было показано, что метод MICE на случайных лесах является наиболее эффективным среди рассматриваемых методов
- Впервые проведен анализ методов заполнения на случайных лесах miceRF, missForest (D.Stekhoven, et al., 2012) не только на наборах данных с пропущенными значениями по типу MCAR, но также и по более сложно реализуемым типам MAR и MNAR

Практическая ценность

- Предложен алгоритм заполнения пропущенных значений в реальных данных
- Реализация метода MICE в Python с привлечением теории марковских цепей

Список литературы

Г.А.Муромцева, et al. (2014) *Распространенность факторов риска неинфекционных заболеваний в российской популяции в 2012-2013гг. Результаты исследования ЭССЕ-РФ*, Кардиоваскулярная терапия и профилактика.

S.V.Buuren (2018) *Flexible Imputation of Missing Data*, 2nd ed., Chapman and Hall/CRC.

D.B.Rubin (1987) *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons Inc.

R.M.Schouten, P.Lugtig, G.Vink (2018) *Generating missing values for simulation purposes: a multivariate amputation procedure*, Journal of Statistical Computation and Simulation.

А.В.Булинский (2017) *Лекции по теории случайных процессов*.

М.Кельберт, Ю.Сухов (2017) *Вероятность и статистика в примерах и задачах*, т. 2, *Марковские цепи как отправная точка теории случайных процессов и их приложения*, Litres.

A.Bendimerad, J.Lijffijt, M.Plantevit, C.Robardet, Tijl de Bie (2020) *Gibbs Sampling Subectively Interesting Tiles*, 18th International Symposium on Intelligent Data Analysis, Germany.