Секция «Теория вероятностей и математическая статистика»

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

Бушмакина Анна Владимировна
Студент (специалист)
Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: anya-merkushina@yandex.ru

Обработка и анализ данных с пропущенными значениями являются актуальным вопросом прикладной математической статистики. Например, выборки, собранные в эпидемиологических исследованиях, зачастую содержат пропущенные значения [1]. Традиционные методы замены отсутствующих значений средним, медианой или константой, а также удаление наблюдений, связанных с утраченными значениями, могут привести к смещенным оценкам и уменьшению объема выборки [2]. По этим причинам в последние десятилетия развились подходы, задача которых — восстановить совместное распределение переменных выборки и подобрать для пропущенного значения наиболее правдоподобное заполнение [3], [4], [5].

Среди современных методов множественная импутация [6] с использованием цепных уравнений (MICE, Multiple Imputation by Chained Equations) [7] отличается своей эффективностью и универсальностью в задачах восстановления отсутствующих данных. MICE основан на семплировании Гиббса [8], широко используемом методе Монте-Карло по схеме марковской цепи [9] для генерации апостериорных распределений. Алгоритм MICE заполняет пропуски итеративно по всем переменным с отсутствущими значениями, используя модель однофакторной импутации, где модель подбирается для каждой переменной по очереди. На каждом шаге пропущенные значения выбираются из условного распределения переменной с учетом наблюдаемых данных и значений других переменных. Множественная импутация предполагает выполнение алгоритма $m > 1$ раз параллельно для получения нескольких заполненных наборов данных. К каждому набору данных применяются стандартные методы анализа, и полученные оценки исследуемых величин объединяются с помощью правил Рубина [6], что позволяет обеспечить более точные оценки по сравнению с традиционными методами импутации.
Цель данной работы заключается в исследовании условий применеия моделей заполнения метода MICE с учетом особенностей данных и потерянных значений. Это позволит обосновать выбор методов вменения пропущенных значений в прикладной математической статистике и обеспечить эффективность метода MICE. Результаты работы включают исследование метода MICE и его эффективности при восстановлении пропущенных значений в наборах данных, сгенерированных по нормальному, равномерному, логнормальному, гамма распределениям, а также в реальных данных. Была проведена оценка точности заполнения пропущенных значений методом MICE для различных моделей импутации (линейная регрессия, случайный лес, k-ближайших соседей), при различных процентах потери данных и типах потери данных (MCAR, MAR, MNAR), а также было выполнено сравнение методов по скорости работы алгоритмов.

Источники и литература

1) 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.

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

3) Y.Duan, Y.Lv, W.Kang, Y.Zhao (2014) *A deep learning based approach for traffic data imputation*, 17th International IEEE Conference on Intelligent Transportation Systems (ITSC).

4) R.Malarvizhi, A.S.Thanamani (2012) *K-nearest neighbor in missing data imputation*, International Journal of Engineering Research and Development.

5) T.H.Lin (2008) *A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data*, Journal Quality and Quantity.

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

7) S.V.Buuren, K.Groothuis-Oudshoorn (2011) *mice: Multivariate Imputation by Chained Equations in R*, Journal of Statistical Software.

8) S.Geman, D.Geman (1984) *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence.

9) C.P.Robert, G.Casella (2004) *Monte Carlo Statistical Methods*, Springer.