Една нова възможност при класификацията на методите за третиране на липсващи данни

гл.ас. Д. Лазаров,
Бургаски свободен университет

Въведение

Класифицирането на обекти в известна степен е скучно занимание. Но в голяма степен то показва отношението и разбирането на автора за обекта. Точно в този дух е настоящата статия. Тя не толкова представя изчерпателен списък на всички съществуващи класификации. Това едва ли е възможно и уместно. Тази статия има за цел по-скоро да покаже „еволюцията“ в разбирането за проблема на липсващите стойности (ЛС). Самият проблем на анализ на ЛС е сравнително нов. До голяма степен той се свързва с работата на Donald Rubin от 70-те години на миналия век и особено откриването на метода на множественото въвеждане. В настоящето обаче той става все по-актуален заради нарастващия обем информация която ползваме и повишеното количество данни които добиваме за да произведем тази информация. ЛС пряко влияят на качеството на информацията. Лесно може да се види, че в зависимост от различния подход към тях може да се достигне до коренно различни изводи и заключения (Enders, C. K., 2010, стр. 13). В настоящата статия се анализират този тип ЛС, които се появяват при отделите значения на изследваните признаци. Те най-често се дължат на липса на отговор на отделния респондент по зададения въпрос или пропуск този отговор да бъде регистриран от изследователския екип. Тук не се разглеждат подходите и класификациите на ЛС при повтарящите се във времето изследвания[1] или ЛС при единиците на изследване, които по някакви причини не са открити или мотивирани да бъдат изследвани.

Познати класификации

Според различните автори тези класификации имат различно съдържание, но могат да се оформят три основни гледни точки. Kalton и Kasprzyk (1986) разглеждат и описват 6 класа на въвеждащи техники по следния начин:

1. Дедуктивни методи: ЛС се въвеждат на базата на логически правила и свързани по дедукция с останалите значения на признаците при конкретната единица (например ако респондента е под 18 годишна възраст то той не може да притежава шофьорска книжка; ако нямате деца то не може да получавате детски).

2. Въвеждане чрез средни стойности или случайни значения: ЛС се заменят със средните стойности на разпределенията (средна аритметична или медиана) или чрез случаен избор на случайно, но наблюдавано значение на целевата променливата.

3. ЛС се заместват със значения от по-стари данни – предходни изследвания, които могат да бъдат аджустирани на базата на някакъв тренд.[2]

4. ЛС се заместват от значения от донорски записи от същата база от данни, които се избират последователно, йерархично или чрез функцията на разстоянието.[3]

5. Въвеждане основано на модел: ЛС се заменят от изчислени значения от регресионен модел или друга непараметрична моделираща техника.

6. Множествено въвеждане: Байсов метод, които въвежда ЛС няколко пъти, което води до множествени бази от данни, които се обединяват в единични оценки посредством правилата на Rubin (Rubin 1987).

Направената класификация е изцяло ориентирана към използваните различни техники за въвеждане на ЛС.

От друга страна Rubin и Little (2002) групират методите за анализ на ЛС в няколко основни категории, които взаимно не се изключват:

Процедури основани на хипотезата за пълната бази данни. Когато даден признак не е

[1] Ако и тези подходи да са по същество същите като описаните.

[2] Cold-deck въвеждане.

[3] Hot-deck въвеждане.

регистриран при някоя единица, решение се търси в „отстраняването" на подобни случаи. Така се анализира винаги една пълна база от данни. В общия случай този подход е най-лесен за изпълнение и при относително малък дял на ЛС той дава задоволителни резултати. Не трябва да се пропуска опасността от сериозно изместване на оценките, като и значително намаляване на тяхната ефективност, особено когато се работи с извадки.

Претеглящи процедури. Резултатите получени от изчерпателни или извадкови изследвания, при които не се наблюдават ЛС, успешно се използват за оценката на тегла за притегляне на резултатите от други изследвания. Теглата обикновено са обратна пропорция на вероятността дадената единица да попадне в извадката. Например, нека yi да бъде значението на признака Y за единицата i в наблюдаваната съвкупност. Средна аритметична за тази съвкупност често се оценя чрез формулата на Horvitz-Thompson (1952 г.):

$$\left(\sum_{i=1}^{n}\pi_i^{-1}y_i\right)\left(\sum_{i=1}^{n}\pi_i^{-1}\right)^{-1}$$

където сумите са от наблюдаваните единици, а π_i е известната априорна вероятност за включване на единицата i в извадката. Претеглящата процедура за липса на отговори променя теглата по начин, сякаш липсващите данни са били част от модела на извадката. Така начина, по който се оценява средната може да се запише по следния начин:

$$\frac{\sum_{i=1}^{n}\left(\left(\pi_i\hat{p}_i\right)^{-1}y_i\right)}{\sum_{i=1}^{n}\left(\pi_i\hat{p}_i\right)^{-1}}$$

където сумите са получени от сумирането на значенията само на отговорилите единици, а $_ip_\frown$ е оценка на вероятността единицата i да даде отговор на изследвания признак. Тази вероятност се получава, обикновено, като дял на отговорилите в извадката.

Процедури основани на въвеждане на стойности. Подходът е ЛС да се въвеждат, така че базата данни да може да се анализира като пълна. Популярна процедура за въвеждане на ЛС е като се използват наблюдаваните значения при дадени единици за заместване на липсващите значения при други. Тук влизат различни подходи като използване на средна стойност получена на базата на наблюдаваните значение за заместване на ЛС или въвеждане на стойности на базата на регресионен модел. С оглед на получаването на реалистични резултати от подобни процедури се налага някои модификации в стандартните анализи, поради различното естество на реалните и въведените данни. Това е свързано с подхода на добавянето на т.н. несигурност във въведените данни и оценките получени на базата на тези данни.

Процедури базирани на модели. Широк клас от процедури основани на дефинирането на модели на базата на наблюдаваните стойности, служещи за основа за заключения чрез максимално-правдоподобните или апостериорните разпределения получени от разглежданите модели. При оценката на параметрите на тези модели най-често се използва метода на максималното правдоподобие. Предимствата на подобен подход е неговата гъвкавост, проверимостта и оценимостта на изследователските допускания и наличието на оценки на дисперсията, в които е отчетен факта на непълнотата на базата данни.

Според Ghosh-Dastidar и Schafer (2003) решения на проблема с ЛС се търсят в три направления, базирани на начина на третиране на базата от данни. Първото направление е да се игнорира проблема и базата данни да се възприеме като пълна. Към тази група решения те причисляват подходи като елиминирането на единиците с ЛС и използването на дъми променливи. Вторият подход е да се представят данните чрез статистически модели като се използват независими променливи свързани с тези, които имат ЛС. Третият подход е да се използва редакция на данните за „изчистването" им, чрез премахване на големите грешки и след това анализиране на редактираните данни като действителни, наблюдавани.

И трите представени класификации са изградени независимо от механизмите на ЛС. Остава впечатлението, че всички методи са еднакво приложими, независимо от действащия механизъм, което може да доведе до объркване.

Механизми на ЛС

Според въведените от Donald Rubin (1987) разграничения на тези механизми те са три: липсващи напълно случайно, липсващи

случайно и не случайно липсващи. При липсващите напълно случайно стойности (ЛНС) появата на самите липсващи стойности може да се разглежда като случайна извадка от единиците в изследваната база от данни. Това означава, че дори и те да бъдат детерминирани от дадена променлива или признак, той не присъства сред наблюдаваните в базата от данни. Вторият по-малко ограничаващ механизъм е липсващи случайно стойности (СЛ). При него появата на липсващи стойности при даден признак е във функция на някои от наблюдаваните променливи, но не и от самия него. Третия и най-проблемен за анализ механизъм е известен като не случайно липсващи (НеСЛ). При този механизъм се появява зависимост между липсващите стойности и самите значения на признака, при който се наблюдават. По друг начин казано, ЛС са във функция на самите себе си.

Една нова възможност за класифициране

От гледна точка на действията, които се предприемат, в зависимост от механизма на ЛС, може да се предложи следната нова класификация.

Елиминационни процедури: Третиране на базата данни, все едно в нея няма ЛС. Единиците, при които де наблюдават липсващи сведения, се отстраняват от анализа. Наблюдават се две вариации на този подход, когато дадена единица с ЛС се отстранява от всички анализи, или когато се отстранява от конкретния анализ, който участва признака с ЛС.

Методи при игнорируеми механизми: Методи, които са приложими при доказване на хипотезата за СЛ или ЛНС механизми.

Методи за въвеждане на ЛС за получаване на пълни бази от данни: В резултат се получават пълни бази от данни. Могат да се направят няколко подгрупи:

- нестохастични методи: Такива методи са методът на независимата средна, регресионното въвеждане, методът „най-близък съсед“, въвеждане по близост на оценените стойности;
- стохастични методи: Стохастична регресия, донорски методи, вариант на въвеждане по близост на оценените стойности;
- многомерни методи като модели със структурни уравнения, непараметрични методи като невронни мрежи и др., които едновременно могат да бъдат стохастични и нестохастични.

Повтарящи се въвеждания: Методи за получаване на обобщени оценки на характеристиките в базата от данни, без реално да се извършва въвеждане на самите ЛС. Такива методи са множественото въвеждане и фракционното въвеждане.

Методи при неигнорируеми механизми: Анализа на ЛС изисква предварително моделиране на механизма на ЛС. Към тази група се отнасят методи като: специфични методи, селекционни модели, смесени модели на поява на ЛС. В резултат, най-често се получават групи от единици, при които механизма на поява на ЛС е игнорируем и могат да се използват методите за въвеждане описани в предходната подточка.

Направената класификация се отнася за липсващите стойности при значенията на признаците в базите от данни при едно изследване и не включва случаите на ЛС поради липса на обхват на съвкупността или ЛС при повтарящи се във времето наблюдения върху една и съща съвкупност.

Заключение

Използването на подобна класификация представя проблема в една нова светлина. Тя акцентира на тясната връзка между действащия механизъм на ЛС и избора на подход за преодоляване на проблемите, произтичащи от ЛС. Това автоматично поставя на преден план необходимостта от точно идентифициране на самите механизми. За съжаление тази част от познанието все още не е достатъчно развита за да може да ни дава нужните отговори във всички възможни ситуации. Все още не е напълно ясно как да се откриват механизмите СЛ и НеСЛ, а те както бе изложено са съпроводени с по-тежки последствия ако не бъдат адекватно анализирани.

Библиография

1. Enders, C. K. (2010) Applied missing data analysis, The Guilford Press
2. Ghosh-dastidar, B. и Schafer, J. L., (2003) Multiple edit/ Multiple imputation for Multivariate Continuous data. Journal of the American Statistical Association, Dec. 2003, Vol.

98, No. 464, Application and Case Studies
3. Horvitz, D.G., Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. Journal of the American Statistical Association. 47, 663-685.
4. Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey data. Survey Methodology 12, 1-16.
5. Little, R.J.A, Rubin, D.B. (1987). Statistical analysis with missing data. New York: Wiley.
6. Little, R.J.A, Rubin, D.B. (2002). Statistical Analysis with Missing Data - 2nd ed., New Jersey: Wiley.
7. Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Survey. New York: Wiley.