# *Използване на невронни мрежи при заместване на липсващи стойности*

*Гл. ас. Д. Лазаров,*
*Бургаски Свободен Университет*

В процеса на събиране на емпирична информация, изследователите често се сблъскват с липса на данни. Те могат да бъдат от различен характер и причинени, като най-вече се дължат на недостатъци в подготовката и методиката на изследването.

В настоящия статия ще разгледаме проблема в контекста на статистическия извадков подход за събиране и обработка на информацията.

В съвременната практика се разглеждат основно три класа липсващи данни:

- Липса на обхват на съвкупността, от която се избира извадката;
- Отказ или невъзможност на единиците, попаднали в извадката да сътрудничат;
- Отказ на единиците, обект на изследване да дадат информация на определени въпроси или загуба или пропуск на подобна информация. Проблемите, възникващи от подобни събития, влияят пряко върху качеството на информацията и точността на оценките. В много случаи, особено при липсващи данни от първите два класа, се прибягва до заместване на единиците с други, които от своя страна имат различни и понякога специфични характеристики. Това води до изместване на основните характеристики на съвкупността. Така получените резултати се превръщат в непредставителни за изучаваната съвкупност [1]. Тези липсващи стойности не означават само по-малка ефективност на оценките, поради редуцирането на размера на базата данни, но също че стандартните методи за анализ на пълни бази данни не могат да бъдат използвани [7]. В случаите на непълни бази данни, рискът от вземане на неправилно решение е изключително висок, защото изкуствено се свиват доверителните интервали, редуцира се статистическата надежност и се получават изместени оценки. Изместването на оценките се дължи и на изчисленията, правени на базата само на получените отговори. Изместването в оценките може да се представи чрез различията между средните значения на наблюдаваните и не наблюдаваните групи от единици:

(1) ,

където $y$е средната величина на изучаваната променлива за цялата съвкупност, 1$y$ е средната стойност на променливата в съвкупността на отговорилите, а 2$y$е неизвестната средна в съвкупността на попадналите единици в изследването, но не отговорили. За оценка на се използва наличната информация от , то изместването на действителната величина може да се представи, чрез:

(2)

Общото изместване или както някои автори го наричат общата грешка, дължаща се на липсата на отговори на респондентите, е функция едновременно на относителния дял на неотговорилите и различието в техните характеристики по изследвания признак. Очевидно е, че при по-голям дял на неотговорилите и малките различията в средните величини на двете групи респонденти се мултиплицират и се отразяват фатално върху крайният резултат от изследването.

**Традиционни подходи при третирането на липсващи стойности**

Един от общопознатите и практикувани методи е известен в литературата като “Listwise delition”, което по същество означава отстраняване от извадката на всички единици с

липсвщи стоийности. При сравнително малък дял на липсващите данни (до 5%) този подход е риемлив, но само в случаите, когато липсите са независими от значенията на променливите, при които се проявяват.

Друг известен подход е т.н. „Pairwise delition". При него се използва цялта налична информация за отделните двойки признаци, които се обработват (корелират) на базата на техните обобщени характеристики, като сума на значенията, средни величини, вариации и ковариации.

Друг подход свързан с анализа на липсващи данни е не тахното отстраняване, а въвеждане (Imputation). В теорията и практиката съществуват различни разновидности на подхода за въвеждане, които в една или друга степен решават задачата. Едни от първите варианти е използването на средната аритметична от значенията на признака вместо липсващата стойност. Познато е също използването на медианната стойност за същата цел. Очевиден е проблемт при работа с категорийни данни като в тези случаи се предлага използването на модалното значение на признака. Тези подходи са съпроводени с натрупване на значителни грешки и отклонения на основните характеристики на признака и основно силно свиване на вариацията, което причинява и значителна промяна на точността на оценките.

Съществува и подходът на заместване на липсващите стойности с наблюдавани значения от извадката при единици със сходни характеристии по останалите признаци.

Методът с значимо влияние сред изследователите в последните години е предложения през 1976 г. от Дейвид Рубин „множествено въвеждане" (Multiple Imputation). Във времето той е получил редица модификации и подобрения, като въвеждането на MCMC (Монте Карло алгоритъм). Основният подходът е да се направят няколко независими оценки на липсващите стойности по даден признак, които после се осредняват за да се плучат значенията, с които да се заместят. Най-често се използват регресионни зависимости за моделиране на процеса на поява на липсващи стойности и тяната оценка. Процетудите се повтарят многократно като на всяка стъпка се оценят обобщени характеристики на базата данни. В края на процеса получените обобщени характеристики се осредняват по оределени правила за да се окончателните оценки елиминиращи влиянието на липсващите данни.

Представения модел работи добре, когато липсващите стойности са случайни или независят от значението на признака, при който се появяват. Има значителни трудности, когато се попадне на случай различен от горепосочените. Определени трудности изникват и когато се обработват категорийни данни, поради честата липса на нормално разпределение при тези признаци и необходимостта от закръгляне на резултатите при процедурата на въвеждане.

**Две решения на проблема за липсващите стойности**

В настоящата работа се съпоставят два метода за анализ на липсващи стойности – **EM – алгоритъм** и **невронни мрежи (ANN)**. Демонстрират се сравнителните предимствата на ANN при анализ на категорийни данни, особено при процент на липсващите стойности е близък до 50.

**Използваната база данни**

В статията е използвана база данни от американския пазар - Employee data.sva, която може да се намери като стандартно приложение в пакета SPSS. В данните има заложени 9 променливи:

- Код (Employee Code)
- Пол (Gender)
- Дата на раждане (Date of Birth)
- Ниво на образование в години (Educational Level (years))
- Позиция на работещия (Employment Category)
- Годишна заплата в момента (Current Salary)
- Стартова годишна заплата (Beginning Salary)
- Период на заемане на съответното място в месеци (Months since Hire)
- Трудов стаж в месеци (Previous Experience (months))
- *Принадлежност към малцинствена група (Minority Classification)*

Обект на изследване е последната променлива, която е представена като категорийна със значения "Да" (Yes) за принадлежност и "Не" (No) в случай, че

единицата на наблюдение не е от малцинствена група. Представени са резултати за 474 единици.

За настоящите ни нужди са направени следните анализи и трансформации. От изходните данни е изчислен регресионен модел със зависима променлива „**Принадлежност към малцинствена група**" и факторни **„Ниво на образование в години", „Позиция на работещия", „Годишна заплата в момента", „Стартова годишна заплата", „Период на заемане на съответното място в месеци", „Трудов стаж в месеци".** Основните показатели, които се анализират са коефициентите на корелация и детерминация на модела и присъщата стандартна грешка – **модел 0** (табл. 1).

**Табл.1**

**Резултати от модела**

| Модел | R | $R^2$ | Аджустиран $R^2$ | Грешка на модела |
|---|---|---|---|---|
| **0** | **0,220(a)** | **0,049** | **0,036** | **0,407** |
| 1 | 0,314(a) | 0,099 | 0,087 | 0,277 |

a) Обясняващи променливи: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Educational Level (years), Employment Category, Current Salary

Резултатите за стойностите на параметрите и техните грешки са поместени в Табл.2.

**Табл. 2**

| Модел | **Коефициенти** | |
|---|---|---|
| | B | Ст. грешка |
| (Constant) | 0,102 | 0,183 |
| Educational Level (years) | 0,003 | 0,009 |
| Employment Category | -0,023 | 0,040 |
| Current Salary | 0,000 | 0,000 |
| Beginning Salary | 0,000 | 0,000 |
| Months since Hire | 0,002 | 0,002 |
| Previous Experience (months) | 0,001 | 0,000 |

От таблицата се вижда, че са коефициентите пред независимите променливи са статистически значими въпреки, че са със стойности много близки до 0. Изключение прави само свободния член в модела.

На следваща стъпка, в полза на експеримента, случайно от зависимата променлива са **премахнати 230 случая**, което представлява **48.52% от информацията**. Резултатите за регресиите са изчислени с помощта на продукта SPSS.

**EM алгоритъм за анализ на липсващи стойности**

В настояшата статия е използван EM алгоритъм (Expectation - Maximization Algorithm). EM е общ итеративен алгоритъм за получаване на максимално правдоподобни оценки в непълни бази данни. В действителност с EM могат да се анализират щирок кръг от проблеми не задължително съдържащи липсваща информация.

В основата на алгоритъма е заложена известната *ad hoc* идея за третиране на липсващите стойности [6]: 1. Замяна на липсващите стойности с техни оценки на базата на наличната пълна информация. 2. Оценка на параметрите на модела на липсващи стойности. 3. Нова оценка на липсващите значения приемайки, че новите параметри са коректно изчислени. 4. Нова оценка на параметрите и т.н. докато се достигне до конвергенция – стабилност в анализираните променливи. На практика в началото се започва с оценката на средната и ковариацията на променливата с липсващи стойности на базата на последователни единични или множествени линейни регресии с останалите променливи в базата данни. Така се получават първоначални оценки за липсващите значения. Следващата M стъпка – максимизиращата – използва първоначално попълнените редове за да оцени отново средните и ковариациите като използва информация за остатъците при регресионните модели. Така получените оценки след тази стъпка се подлагат отново на регресионна оценка – E стъпка и т.н.

Резултатите от прилагането на анализа и

провеждането на регресионния анализ по отношение на дефинираната зависима променлива са поместени в табл. 1, модел 1. Специфично те могат да се видят и в табл 3.

**Табл. 3**

**Резултати от модела EM**

| Модел | R | $R^2$ | Аджустиран $R^2$ | Грешка на модела |
|---|---|---|---|---|
| 1 | 0,314(a) | 0,099 | 0,087 | 0,277 |

а) Обясняващи променливи: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Educational Level (years), Employment Category, Current Salary

От резултатите се вижда че, коефициента на корелация и детерминация е изместен значително. В оригиналните данни стойностите са значително по-малки отколкото в оценените. Това е придружено и от подценяване на грешката на модела, което води до изкуствено увеличаване на точността на оценките, един от проблемите при анализа на липсващи стойности. Това може да се види в следващата табл. 4 показваща стойността на коефициентите на регресията и тяхната точност.

**Табл. 4**

| Модел | Коефициенти | |
|---|---|---|
| | B | Ст. грешка |
| (Constant) | 0,114 | 0,125 |
| Educational Level (years) | 0,011 | 0,006 |
| Employment Category | -0,029 | 0,028 |
| Current Salary | 0,000 | 0,000 |
| Beginning Salary | 0,000 | 0,000 |
| Months since Hire | 0,000 | 0,001 |
| Previous Experience (months) | 0,001 | 0,000 |

**Невронна мрежа (ANN)**

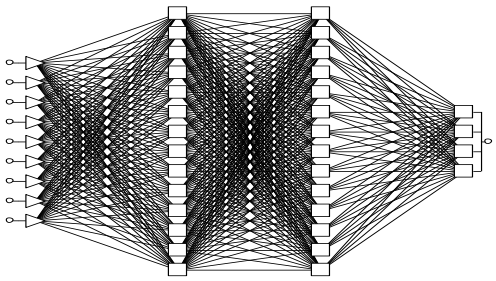
Поради естеството на данните с липсващи стойности - една категорийна променлива на номинална скала бе избрана за да се моделира връзката аналогично на заложеното в EM алгоритъма. Така като резултативна, в невронния модел, се дефинира отново променливата „**Принадлежност към малцинствена група”,** която приема две категорийни стойности “Да” и „Не”. Избраната мрежа е Многослоен персептрон (MPL) с 9 входни променливи, 28 скрити неврона в 2 междинни слоя и изходен слой. Мрежата е обучена по методите Back propagation и Conjugate gradient descent [4] в специфично съчетание. Резултатите за мрежата могат да се видят в табл 5.

**Табл. 5**

| Вид | Грешка | Входни неврони | Скрити неврони | Ефективност |
|---|---|---|---|---|
| MLP | 0.1598457 | 6 | 28 | 0.8983051 |

Графичния вид на мрежата може да се види на диаграма 1.

**Диаграма 1**



За обучение на мрежата са използвани само редовете от базата данни, в които няма липсващи стойности в зависимата променлива, т.е. обучаващия набор се състои от 237 наблюдения, за контролен 118, а за тестов набор 119. След обучението цялата база данни (редове с и без липсващи стойности) се оценя с помощта на мрежата. Така се получават и оценки за липсващите стойности в зависимата променлива. Получените резултати са в категорийна форма и не изискват прехода от метрирани към неметрирани данни. Мрежата е изградена и обучена с помощта на пакета Pythia.

Така получените резултати са включени в регресионния модел като зависима променлива и резултатите са поместени в табл. 6, модел 2.

**Табл. 6**
**Резултати от модела ANN**

| Модел | R | $R^2$ | Аджустиран $R^2$ | Грешка на модела |
|---|---|---|---|---|
| 2 | 0,241(a) | 0,058 | 0,046 | 0,223 |

a) Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Educational Level (years), Employment Category, Current Salary

Като се вижда, използването на невронна мрежа е значително предимство по отношение на зависимата променлива. Почти се елиминира изместването на коефициентите на корелация и детерминация, но остава проблема с подценяването на стандартната грешка. В табл. 7 са поместени резултатите от регресионни модел при използване на попълнените данни за *Принадлежност към малцинствена група (Minority Classification)* получени от мрежата.

**Табл. 7**

| Model | **Коефициенти** | |
|---|---|---|
| | B | Ст. грешка |
| (Constant) | -0,360 | 0,100 |
| Educational Level (years) | 0,014 | 0,005 |
| Employment Category | 0,062 | 0,022 |
| Current Salary | 0,000 | 0,000 |
| Beginning Salary | 0,000 | 0,000 |
| Months since Hire | 0,003 | 0,001 |
| Previous Experience (months) | 0,000 | 0,000 |

**Заключение**

Проведеното изследване показва, че може да се търси решение на проблемите с липсващи стойности чрез използване на невронните мрежи за моделиране на връзките. Сами по себе си те се оказват по добър инструмент за тази цел в сравнение с регресионните модели, които са зависими от емпиричното разпределение на променливите, включени в тях. От друга страна, съчетанието на известните алгоритми за анализ на липсващи стойности с интегрирани в тях невронни модули би подобрило решенията на задачата за липсващите данни. Не трябва да се забравя, че в настоящата работа изкувствено бяха получени липсващи стойности, които са по модел ЛНС (липсващи напълно случайно) по отношение на включените в анализа променливи. Интерес представлява анализ на данни, които са с модел ЛС (липсващи случайно) и НСЛ (не случайно липсващи). Интересен и друг аспект на използването на невронните мрежи свързан с априорна диагностика на модела на липсващи стойности, както и с апостериорна оценка на ефективността на приложените въвеждащи процедури. Невронните мрежи са мощен класификационен инструмент и се очаква да се получат добри резултати и в комбинация с други *hot dec* методи за въвеждане на липсващи данни.

**Литература**

1.Богданов, Б, (1988). "Измерване изменението на оценките от наблюдението на домакинските бюджети", Социологически преглед, бр.3.

2.Allison, P.D. (2002). Missing Data. Sage University Papers Series on Quantitative Applications in Social Science, 07-136. Thousand Oaks, CA: Sage.

3.Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. Sociological Methods and Research 28: 301-309.

4.Bishop, C. M., (1995). Neural network for pattern recognition. Oxford UP.