МЕТОДЫ ВОССТАНОВЛЕНИЯ НЕПОЛНЫХ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ АНАЛИЗА

Судаков Б. Д., Герман О. В.
Кафедра информационных технологий автоматизированных систем,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: sudakov.bogdan666@gmail.com, ovgerman@bsuir.by

В данной работе рассматриваются методы восстановления неполных данных, которые позволяют повысить точность анализа и улучшить качество получаемых результатов. Представлен обзор различных подходов, включая регрессионный анализ, метод хот-дек и алгоритм ZET. Обсуждается применимость каждого метода в зависимости от специфики задачи и характера неполноты данных.

Введение

При проведении статистического анализа на практике ограничиваются анализом не всей генеральной совокупности в целом, а лишь некоторого выборочного числа наблюдений. Анализируемая выборка должна отвечать критериям качественности и полноты. В реальности приходится сталкиваться с ситуацией, когда некоторые из свойств одного или нескольких объектов отсутствуют – возникает ситуация данных с пропусками, что значительно осложняет математическую обработку, так как смещение основных статистических характеристик, таких как математическое ожидание или дисперсия, например, возрастает прямо пропорционально числу пропусков. На сегодняшний день в математической статистике существует несколько путей решения проблемы неполных данных:

- Исключение некомплектных объектов из исходной выборки;
- Применение специально разработанных математических методов анализа неполных данных, таких как метод взвешивания или метод максимального правдоподобия и EM-алгоритм [1];
- Восстановление пропусков (наиболее распространены методы заполнения по среднему и по регрессии).

I. Методы восстановления пропусков

Согласно классификации Литтла и Рубина [2], в данных могут встречаться:

1. Полностью случайные пропуски (missing completely at random (MCAR));
2. Случайные пропуски (missing at random (MAR));
3. Неигнорируемые пропуски (non-ignorable missingness).

С полностью случайными пропусками (MCAR) мы имеем дело тогда, когда пропуски случайно распределены в массиве данных по всем переменным. Наличие MCAR можно проверить статистически t-тестом или хи-квадрат тестом. В модуле SPSS Missing Values Analysis (MVA) есть опция Little's MCAR test, которая на базе статистики хи–квадрат проверяет данные на MCAR. Если в тесте наблюдается незначимый уровень критерия, то мы имеем дело с полностью случайными пропусками данных. Случайные пропуски (MAR) данных встречаются тогда, когда пропуски в массиве данных случайно распределены не по всем переменным, а только внутри каких-либо определенных подгрупп переменных. Такое распределение пропусков в данных случается гораздо чаще, чем MCAR.

Для борьбы с этими двумя видами пропусков применяют восемь основных классов методов:

1. Анализ полных наблюдений (listwise deletion);
2. Методы, использующие доступную информацию (pairwise deletion);
3. Подстановка среднего по выборке (mean substitution);
4. Метод хот-дек (hot deck);
5. Регрессионный анализ (regression);
6. Оценка с помощью максимизации правдоподобия (maximum likelihood estimation);
7. Подстановка с помощью факторного анализа (factor analysis substitution);
8. Модель множественного восстановления данных (multiple imputations method).

Два первых метода широко распространены в исследовательской практике.

Методы с 3-го по 7-ой используют принцип однократной подстановки восстановленных тем или иным способом данных и могут использоваться, если пропуски распределены случайно (MAR). В основе методов лежит принцип вычисления и подстановки взамен каждого пропущенного значения одного нового значения. Кратко рассмотрим эти методы.

Метод хот-дек (hot deck) представляет собой метод подстановки среднего по выборке с некоторым количеством модификаций. Наиболее простой вариант — сортировка респондентов по ключевым переменным, тогда респонденты со схожими ответами находятся рядом друг с другом [3]. В качестве ключевых чаще всего выступают социально-демографические переменные, но можно использовать и другие переменные, имеющие корреляции с переменной с пропущенными

данными. При восстановлении пропущенные значения переменной заимствуются из предыдущего наблюдения.

Регрессионный анализ. В зависимости от типа данных используются либо множественная линейная, либо логистическая регрессия. На базе наблюдений, не содержащих пропущенных данных, вычисляются коэффициенты регрессии, и далее с их помощью восстанавливается пропущенное значение зависимой переменной. Если не говорить об обычных проблемах регрессии, таких как мультиколлинеарность, гомоскедастичность и т.д., то можно обозначить две проблемы, связанные с её использованием для восстановления пропущенных данных. Во-первых, из-за самой природы регрессии мы полностью исключаем случайные вариации. Это ведет к тому, что при большой доле пропущенных значений становится очень заметным смещение результатов по направлению к средним оценкам. Для борьбы с этим используется метод случайной подстановки, при котором к вычисленному значению прибавляются случайные величины. Во-вторых, используя в уравнение регрессии слишком большой набор независимых переменных, мы рискуем моделировать шум вместо каких-то осмысленных значений переменных.

Алгоритм ZET является детально проработанной и апробированной технологией верификации экспериментальных данных, основанной на гипотезе их избыточности. Главная идея алгоритма ZET заключается в подборе «компетентной матрицы», используя данные из нее находят параметры зависимости, которая применяется для прогнозирования пропущенного значения. Субъективизм определения размерности «компетентной матрицы» приводит к учету неинформативных и шумовых факторов и смещению оценки неизвестного значения. Основное отличие алгоритма состоит в определении оптимального размера «компетентной матрицы». Рассмотрим подробнее методику алгоритма ZET.

II. Алгоритм ZET

В основе алгоритма ZET лежат три предположения. Первое (гипотеза избыточности) состоит в том, что реальные таблицы имеют избыточность, проявляющуюся в наличии похожих между собой объектов (строк) и зависящих друг от друга свойств (столбцов). Если же избыточность отсутствует (как, например, в таблице случайных чисел), то предпочесть один прогноз другому невозможно. Второе предположение (гипотеза локальной компактности) состоит в утверждении, что для предсказания пропущенного элемента нужно использовать не всю таблицу, а лишь ее «компетентную» часть, состоящую из элементов строк, похожих на строку i, и элементов столбцов, похожих на столбец j. Остальные строки и столбцы для данного элемента неинформативны. Их использование лишь разрушало бы локальную компактность подмножества компетентных элементов и ухудшало точность предсказания. Третье предположение (гипотеза линейных зависимостей) заключается в том, что из всех возможных видов зависимостей между столбцами (строками) в алгоритме ZET используются только линейные зависимости. Если зависимости носят более сложный характер, то для их надежного обнаружения требуется такой большой объем данных, который в реальных задачах встречается нечасто. Для различных прикладных задач были сделаны многочисленные модификации базового алгоритма ZET, отличающиеся своим назначением и наборами разных режимов работы. Программы заполнения пробелов могут работать в одном из следующих режимов:

- Заполнение всех пробелов;
- Заполнение только тех пробелов, ожидаемая ошибка для которых не превышает заданной величины;
- Заполнение пробелов только на базе информации, имеющейся в исходной таблице;
- Заполнение каждого следующего пробела с использованием исходной информации и прогнозных значений ранее заполненных пробелов.

III. Выводы

В данной работе рассмотрены методы восстановления неполных данных, которые позволяют повысить точность анализа и улучшить качество принимаемых решений. Применение алгоритмов восстановления неполных данных позволяет избежать потери важной информации, которая может негативно повлиять на точность анализа. Кроме того, использование этих алгоритмов помогает снизить риск ошибок, связанных с пропуском важных данных. Таким образом, алгоритмы восстановления неполных данных являются важным инструментом для повышения точности и надежности анализа данных, а также обеспечения более качественного принятия решений на основе полученных результатов.

Список литературы

1. Rubin, D. B. Statistical Analysis with Missing Data / D. B. Rubin, R. J. A. Little // New York: John Wiley & Sons – 2017. –P. 40–43.
2. Wedel, M. Factor Analysis and Missing Data / M. Wedel, W. A. Kamkura // Journal of Marketing Research. – 2000. – № 4. – С. 490–498.
3. Косьяненько, А. В. Опыт восстановления пропущенной рыночной информации на основе Байесовского подхода / А. В. Косьяненко // Издательство: Инфра-М, 2016. – 193 с.