УДК 51(083)

МЕТОДИКА ВОССТАНОВЛЕНИЯ ДАННЫХ С ПРОПУСКАМИ

Варламов М.С.
Научный руководитель канд. техн. наук Даничев А.А.
Сибирский Федеральный Университет

Введение

Проблема пропущенных значений достаточно актуальна, к примеру для социологии. Причинами неполноты данных опроса могут служить множество факторов: невнимательность респондента, ошибки в анкете, различие в данных анкет (в нашем случае) и т.д. В результате на этапе анализа данных мы имеем неполный массив. Именно для таких случаев и предназначены методы восстановления данных с пропусками.

Обзор современных методов восстановления данных с пропусками[1]

На данный момент существует множество методик позволяющих эффективно восстанавливать данные с пропусками, но у каждой из них есть свои плюсы и минусы. Данные методики в большинстве своем используются при незначительном количестве пропусков. Перечислим наиболее распространенные методы с указанием их основных особенностей:

1 Исключение строк с наличием пропусков. Данный метод легко реализуем, но необходимым условием его применения является следование данных требованию MCAR (missing completely at random), т.е. пропуски в данных по переменным должны быть полностью случайными. Кроме того, он обычно применяется лишь при незначительном количестве пропусков в таблице, иначе полученная на выходе таблица данных становится непредставительной. Главный недостаток такого подхода обусловлен потерей информации при исключении неполных данных.

2 Заполнение пропусков средними по столбцу значениями. Данный метод также легко реализуем, но его применение имеет смысл только в случае удовлетворения данных условию MAR (missing at random), т.е. когда пропуски в данных по переменным являются случайными и сам механизм пропусков несущественен. К недостаткам метода относят вносимые искажения в распределения данных, уменьшение дисперсии.

3 Метод ближайших соседей. В основе метода лежит механизм поиска строк таблицы, которые по определенному критерию являются ближайшими к строке с пропусками. Для заполнения пропуска значения данной переменной (в фиксированном столбце) у соседних строк усредняются с определенными весовыми коэффициентами, обратно пропорциональными расстоянию к строке с пропуском. При большом количестве пропусков данный метод также практически неприменим, поскольку базируется на существовании связей между строками в таблице.

---

[1] На основе материала из книги Загоруйко Н.Г. Методы распознавания и их применение. – М.: Советское Радио, 1972.

4 Метод сплайн-интерполяции. Для успешного применения необходимо, чтобы данные следовали условию MAR. Недостатки метода следуют из самой его идеи. Например, в случае восстановления группы пропусков, следующих подряд друг за другом, результат аппроксимации сплайном данной группы не всегда может дать оценки, приближающиеся с достаточной точностью к значениям, которые могли бы быть на месте пропусков.

5 Метод максимальной правдоподобности и EM-алгоритм. Метод требует проверки гипотез о распределении значений переменных. Применение осложняется при большом количестве пропущенных значений переменной. Особенность данного метода состоит в построении модели порождения пропусков с последующим получением выводов на основании функции правдоподобия, построенной при условии справедливости данной модели, с оцениванием параметров методами типа максимального правдоподобия. Отметим, что для данных методов возможно построение моделей, учитывающих конкретную специфику области, и, как следствие, возможна постановка более слабых условий к данным (слабее MAR).

6 Алгоритмы ZET и ZetBraid. По сути, 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 ZetBraid состоит в определении оптимального размера «компетентной матрицы». Данные алгоритмы хорошо показали себя, но статистическая оценка неизвестного значения исключительно на основе корреляционно регрессионного анализа и необходимость задания ряда важных параметров приводит к необходимости убедиться в правдоподобности восстановленных значений.

Рассмотрим подробнее методику алгоритма ZET.

Алгоритм ZET[2]

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 не возможно. Второе предположение (гипотеза локальной компактности) состоит в утверждении, что для предсказания пропущенного элемента b_{ij} 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

[2] Материал из книги Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999.

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:

1 Заполнение всех пробелов.

2 Заполнение только тех пробелов, ожидаемая ошибка для которых не превышает заданной величины.

3 Заполнение пробелов только на базе информации, имеющейся в исходной таблице.

4 Заполнение каждого следующего пробела с использованием исходной информации и прогнозных значений ранее заполненных пробелов.

Заключение

При анкетировании в основном встречаются 2 системы оценки ответов: бальная и ранжирование. Для каждой из этих систем оптимальны определенные методы заполнения пропусков. Например при бальной системе оценок удобно применять бинарные , вещественнее шкалы. Тогда как при ранжировании удобно применять уже совершенно другие алгоритмы. Исходя из данных полученных при заполнении пропусков с помощью алгоритма ZET, можно сказать, что данный алгоритм весьма эффективен при прогнозировании большого объема однотипных данных с большим числом пропуском. Так же стоит заметить, что данный алгоритм эффективен при любом выборе метода оценки анкетирования, будь то ранжирование или бальная система. Что дает возможность точнее оценивать результаты анкетирования студентов и позволяет использовать полученные им данные для определения значимости образовательных предметов при поступлении на работу выпускников Вузов.

Список литературы

1. Загоруйко Н.Г. Методы распознавания и их применение. – М.: Советское Радио, 1972.
2. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modeling & New Technologies.; Vol. 6.2004.; Стр.55 – 56.
3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999.