Мартышенко Н.С., Мартышенко С.Н. Методы восстановления данных анкетных опросов // Системы управления и информационные технологии, 2007, 3.3.(29). – С 364-368.

2007

УДК 519.254

С.Н. Мартышенко, Н.С. Мартышенко

МЕТОДЫ ВОССТАНОВЛЕНИЯ ДАННЫХ АНКЕТНЫХ ОПРОСОВ

S.N. Martishenko, N.S. Martishenko

THE METHODS OF QUESTIONNAIRE SURVEY DATA RECOVERY

Данные анкетных опросов часто содержат ошибки, которые могут достигать достаточно высокого уровня и содержать пропуски. В работе рассматривается метод восстановления данных, основанный на использовании алгоритма распознавания многомерных статистических данных. Рассмотрены методы оценки точности восстановления данных.

Questionnaire survey data often contain multiple mistakes and gaps. The paper deals with the method of questionnaire survey data recovery. It is based on using a recognition algorithm for multi-dimensional statistical date base.

В последние годы в исследовании социально-экономических процессов все большее распространение получает анкетный опрос. Однако большинство анкетных опросов ограничивается небольшими выборками, редко превышающими нескольких сот элементов. Многие исследователи вообще опрашивают не более ста респондентов. Такие исследования могут служить только для выработки предварительных гипотез или апробации вопросов анкеты с точки зрения их восприятия респондентами. Даже для изучения связи между какими-либо двумя признаками такого объема выборки часто недостаточно. Использование малых выборок, с одной стороны, может быть связано с ограниченными ресурсами, с другой стороны, многие исследователи и не ставят перед собой сложных задач, требующих привлечения наукоемких технологий. При этом и ценность результатов не очень высока. Наибольшую ценность имеют данные, которые собираются с целью использования их в экономико-математических моделях, применимых для выработки управленческих решений. При разработке таких моделей данные анкетных опросов могут быть использованы для выявления и описания структуры социально-экономических групп, а также оценки некоторых параметров моделей. В этом случае могут потребоваться выборки, содержащие до тысячи и более объектов. В отдельных случаях необходимо иметь данные, полученные в результате нескольких опросов, относящихся к различным временным интервалам (мониторинг ситуации). Поскольку модели требуют не только качественной формулировки гипотез, но и количественных оценок, возникает проблема оценки достоверности данных и характеристик по ним рассчитанным.

Для большинства данных, полученных в результате опросов характерно наличие большого количества грубых ошибок [1], которые приводят не только к искажению оценок, а часто вообще исключают возможность использования отдельных методов анализа данных. Поэтому, прежде чем

приступить к содержательному анализу данных, целесообразно исследовать качество первичных данных. Такой анализ оказывается полезным при проведении систематических анкетных опросов. В этом случае исследователя интересуют не только сами ошибки, но и причины их возникновения с тем, чтобы в последующем постараться так организовать сбор данных, чтобы снизить уровень ошибок.

При разработке методов анализа анкетных данных каждый объект выборки удобно интерпретировать как значение признака многомерной выборки.

Грубые ошибки в данных особенно вредны при совместном анализе нескольких признаков многомерных данных, когда нецелесообразно отбрасывать все многомерное наблюдение из-за ошибки в одном из признаков. Аналогичная ситуация возникает и в случае отсутствия данных в одном из признаков многомерного наблюдения.

Повысить качество многомерных данных можно путем применения алгоритмов восстановления данных. В данной работе мы рассмотрим один из методов восстановления данных. При этом будем считать, что грубые ошибки или выбросы в данных были выявлены на предыдущих этапах обработки данных. Методы решения задачи выявления грубых ошибок были рассмотрены в одной из работ авторов [2].

Преимуществом предлагаемого алгоритма является то, что он может быть применим к смешанным признакам, то есть признаки многомерного наблюдения могут быть измерены в различных измерительных шкалах (номинальной, порядковой, относительной). Алгоритм был разработан на основе идей метода интегральной диагностики, опубликованного в работе [3]. Этот алгоритм относится к классу алгоритмов распознавания образов.

При восстановлении данных грубые ошибки или выбросы мы будем рассматривать, как ситуацию отсутствия данных. Работу алгоритма рассмотрим по этапам.

Первый этап: выбор признаков. Предположим, что из всех признаков выборки данных анкетного опроса выбран признак, для которого мы хотим восстановить ряд отсутствующих значений. Обозначим этот признак, как Y - восстанавливаемый или прогнозируемый признак. В реальной ситуации количество пропущенных значений редко превышает 5-7% от общего количества наблюдений. Если таких данных более 10% целесообразно проанализировать причину возникновения пропуска. Возможно, она кроется в самой постановке вопроса анкеты.

Для восстановления значений отберем ряд признаков, относительно которых мы предполагаем, что они в совокупности должны в какой-то степени оказывать влияние на значения признака Y. Такие признаки будем называть восстанавливающими или прогнозирующими: $X_1, X_2, \ldots, X_j, \ldots, X_m \quad (j = \overline{1,m})$. Считается, что для всех значений признака Y (в том числе и пропущенных) определены значения восстанавливающих

признаков. Будем полагать, что восстанавливаемый признак может принимать k возможных различных значений, а каждый восстанавливающий признак $s_j\,(j=\overline{1,m})$ возможных значений. Для простоты предположим, что прогнозируемый признак Y может принимать значения из ряда целых чисел (1,2, …, k). Такое предположение вполне согласуется с признаками, представленными в номинальной или порядковой шкале. Если какой-либо признак был измерен в относительной шкале, то он должен быть предварительно преобразован к порядковой шкале путем ранжирования.

Второй этап: формирование обучающей и рабочей выборок. Пусть вся выборка состоит из n наблюдений. Разобьем выборку на две. Первая, из n_1 наблюдений имеют все значения и восстанавливаемого признака Y и восстанавливающих $X_1, X_2, \ldots, X_j, \ldots, X_m$, Оставшиеся n_2 наблюдения имеют значения только восстанавливающих признаков, а значения восстанавливаемого признака отсутствуют. Задача состоит в восстановлении n_2 значений признака Y на основе информации содержащейся в обоих выборках. Первую выборку из n_1 объектов можно рассматривать как обучающую выборку, вторую выборку из n_2 объектов, как рабочую выборку. В обучающей выборке признак Y служит для задания разбиения объектов на классы. В обучающей выборке каждому многомерному наблюдению восстанавливающих признаков $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}\ (i=\overline{1,n_1})$ поставлен в соответствие номер класса - y_i, принимающий одно из возможных значений признака Y. То есть, обучающая выборка состоит из k классов Подсчитаем объекты с одним и тем же значением y_i и получим объемы классов - $\pi_r\ (r=\overline{1,k})$:

$$\sum_{r=1}^{k}\pi_r = n_1 \qquad (1)$$

Для элементов рабочей выборки номера классов не определены, так как для них неизвестны значения прогнозируемого признака.

Третий этап: расчет эталона классов. Наблюдения обучающей выборки используем для построения эталона классов, который затем будут применяться для распознавания элементов рабочей выборки с целью определения номеров классов, что эквивалентно определению для них значений прогнозируемого признака Y. Рассмотрим правило построения эталона более подробно.

Эталон состоит из k элементов - по числу классов. Расчет элементов эталона классов рассмотрим на примере одного признаку - X_j. Остальные элементы рассчитываются аналогично.

Преобразуем каждый прогнозирующий признак обучающей выборки $X_j\ (j=\overline{1,m})$ к двоичному виду. Каждый такой признак будет представлен

s_j $(j=\overline{1,m})$ двоичными признаками по числу возможных значений признака X_j.

Таким образом, каждое значение j –го признака обучающей выборки будет представлено набором значений бинарных признаков $\eta_i^j=(\xi_{i1}^j,\xi_{i2}^j,\ldots,\xi_{it}^j\ldots,\xi_{is_j}^j)$, где $i=\overline{1,n_1}$, $j=\overline{1,m}$, $t=\overline{1,s_j}$ для каждого из которых известен номер класса y_i $(i=\overline{1,n_1})$. Для бинарных значений справедливо:

$$\sum_{t=1}^{s_j}\xi_{it}^j=1, \tag{2}$$

$$\sum_{i=1}^{n_1}\sum_{t=1}^{s_j}\xi_{it}^j=n_1 \tag{3}$$

Сложим вектора η_i^j, относящиеся к одному классу. В результате, получим матрицу $\tilde{\Omega}_j$ размерности $r\times s_j$ $(r=\overline{1,k})$. Каждая строка матрицы представляет собой частотный ряд бинарного разложения признака X_j. Разделив построчно элементы матрицы $\tilde{\Omega}_j$ на число элементов класса π_r $(r=\overline{1,k})$, получим k относительных частотных рядов бинарного разложения признака X_j. Преобразованную матрицу обозначим Ω_j

$$\Omega_j=\begin{pmatrix}\omega_{11}^j & \omega_{12}^j & \ldots & \omega_{1t}^j & \ldots & \omega_{1s_j}^j\\ \omega_{21}^j & \omega_{22}^j & \ldots & \omega_{2t}^j & \ldots & \omega_{2s_j}^j\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \omega_{r1}^j & \omega_{r1}^j & \ldots & \omega_{rt}^j & \ldots & \omega_{rs_j}^j\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \omega_{k1}^j & \omega_{k1}^j & \ldots & \omega_{kt}^j & \ldots & \omega_{ks_j}^j\end{pmatrix} \tag{4}$$

По данным матрицы Ω_j сформируем элемент эталона классов по j – ому признаку:

$$V_j=(v_1^j,v_2^j,\ldots,v_t^j,\ldots,v_{s_j}^j). \tag{5}$$

Значения параметров v_t^j определяется по номеру строки матрицы Ω_j (ассоциированной с номером класса) при котором достигается максимум столбца t $(t=\overline{1,s_j})$. То есть параметры v_t^j принимают значения из ряда натуральных чисел 1,2,…,k. Аналогично рассчитываются все остальные элементы эталона по всем прогнозирующим признакам. В результате получим эталон классов, заданный совокупностью векторов V_j $(j=\overline{1,m})$. Общее количество параметров эталона φ:

$$\varphi = \sum_{j=1}^{m} s_j \ . \qquad (6)$$

Четвертый этап: преобразование наблюдений рабочей выборки. Данные рабочей выборки имеют ту же структуру, что данные обучающей выборки за исключением номеров классов, которые для объектов рабочей выборки не известны. Преобразуем прогнозирующие признаки рабочей выборки к бинарному виду. Тогда каждое наблюдение рабочей выборки будет представлено *m* векторами:

$$W_\tau^j = (w_{\tau 1}^j, w_{\tau 2}^j, \ldots, w_{\tau t}^j, \ldots, w_{\tau s_j}^j), \ \ j = \overline{1,m}; \tau = \overline{1,n_2} \ . \qquad (7)$$

Размерность векторов совпадает с размерностью аналогичных векторов эталона классов V_j ($j = \overline{1,m}$). В отличии от эталона вектора W_τ^j принимают бинарные значения (то есть 0 или 1).

Пятый этап: восстановление номеров классов для рабочей выборки. Эта процедура эквивалентна восстановлению отсутствующих значений прогнозируемого признака *Y*, так как номер класса ассоциируется с одним из возможных значений признак *Y*.

Определение номера класса (распознавание) для элементов рабочей выборки производится при сопоставлении элементов эталона классов V_j с векторами, описывающими наблюдения рабочей выборки W_τ^j. В результате сопоставления векторов образуется *k* сумм Θ_r ($r = \overline{1,k}$) (по числу классов). Расчет сумм производится по правилу: если параметр $w_{\tau t}^j = 1$ ($j = \overline{1,m}; t = 1, s_j; \tau = \overline{1,n_2}$), то сумма с номером v_t^j получает приращение равное единице. Сумма, получившая наибольшее значение определяет номер класса рабочей выборки или значение прогнозируемого признака. В случае, если максимальных значений несколько, то предпочтение отдается классу с большим объемом обучающей выборки π_r ($r = \overline{1,k}$).

Накопление суммы по единице целесообразно при одинаковом количестве градаций прогнозирующих признаков $s_j = s$ ($j = \overline{1,m}$). При различном количестве градаций формируются взвешенные суммы. В сумму класса добавляется не единица, а вес класса:

$$Q_j = \frac{s_j}{\sum_{j=1}^{m} s_j}. \qquad (8)$$

Резюме по алгоритму. Алгоритм допускает несколько модификаций. В любом случае результатом является прогноз признака *Y* для наблюдений рабочей выборки. Возникают два вопроса. Первый – “какова точность восстановления ?”. То есть, в какой степени исследователь может доверять полученному результату. Очевидно точность зависит не только

от вида используемого алгоритма, но и того какие восстанавливающие признаки были использованы для прогнозирования признака *Y*.

Для оценки точности восстановления можно использовать два способа.

Первый способ основан на использовании бустреп-технологии [4], которая в последнее время находит все большее распространение. Суть подхода состоит в том, что из обучающей выборки случайным образом изымается часть наблюдений для которых"временно забываются" значения прогнозируемого признака *Y*, которые затем восстанавливаются с помощью эталона. По восстановленным данным рассчитывается матрица ошибок распознавания *A*, в которой указывается количество правильно и ошибочно восстановленных значений признака *Y*. Элемент матрицы ошибок a_{gf} - определяет количество элементов класса g отнесенных к классу f $(g=\overline{1,k};\ f=\overline{1,k})$. После нормализации матрицы *A* по строкам получим матрицу относительных частот ошибок *P*. Тогда для характеристики качества восстановления данных можно использовать оценку средней вероятности ошибки:

$$p_{cp}=\sum_{g=1}^{k}\sum_{f=1}^{k}\frac{p_{gf}}{(k^2-k)} \tag{9}$$

При многократном запуске случайного механизма можно оценить даже доверительный интервал оценки средней вероятности ошибки. При желании по матрице P можно оценить ошибки первого и второго рода, но как правило в таких задачах, этого не требуется. Особенность этого метода состоит в том, что в обучении используются все n_1 объектов обучающей выборки.

Второй способ основан на использовании процедуры скользящего экзамена. В этом случае, из обучающей выборки поочередно изымаются по одному наблюдению, а остальные (n_1-1) наблюдений используются для построения эталона классов, с помощью которого восстанавливается значение прогнозируемого признака для изъятого наблюдения. После распознавания изъятое наблюдение возвращается в выборку и изымается следующее и так далее. Полученная ошибка является верхней оценкой действительной ошибки. Однако при больших выборках оба способа дают близкие результаты. В первом способе эталон классов рассчитывается только один раз, а во втором эталон необходимо рассчитать n_1 раз. Поэтому скорость работы первого алгоритма на много выше чем второго.

В задаче восстановления пропущенных значений уровень ошибки до 10% можно считать очень хорошим. В отдельных случаях допустима ошибка до 20% и даже 25%. Настолько высокие допуски уровня ошибки объясняются тем, что количество восстанавливаемых значений чаще всего

не превосходит 5% от общего объема выборки и их вклад в конечный результат незначителен.

Со вторым вопросом – выбора прогнозирующих признаков пока нет полной ясности. Разработка формализованных методов выбора таких методов находится в настоящее время на стадии исследования. На сегодняшний день можно рекомендовать для отбора прогнозирующих признаков оценивать зависимость между каждым таким признаком и восстанавливаемым признаком с помощью критерия χ^2. Очевидно, если критерий показывает, что прогнозирующий и прогнозируемый признак независимы, то использовать такой прогнозирующий признак не целесообразно. При выборе прогнозирующего признака могут быть полезны и качественные рассуждения.

Алгоритм был реализован как дополнительный модуль EXCEL. Этот модуль был включен в состав специализированного комплекса программных средств по обработке анкетных данных[2].

Исследование возможностей алгоритма было произведено на множестве примеров модельных данных многомерных нормальных выборок. Зависимость признаков при моделировании выборок задавалась корреляционными матрицами [5]. Апробация алгоритма была произведена на примере данных анкетных опросов, проводимых в ходе маркетинговых исследований туристского комплекса региона.

Во многих случаях алгоритм высокую эффективность работы. Но, как и любой статистический метод, предложенный метод восстановления данных нельзя признать абсолютным инструментом. Применение такого способа восстановления целесообразно использовать совместно с другими способами. Часто очень эффективным средством является логический способ восстановления данных. Такие алгоритмы также вошли в состав разработанного нами комплекса программных средств. [1].

Практическая значимость рассмотренного алгоритма наиболее полно проявляется при совместном использовании множества процедур обработки анкетных данных, входящих в разработанный программных средств, который обеспечивает высокую технологичность обработки анкетных данных при систематических исследованиях социально-экономических процессов.

Литература

1. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Многомерные статистические методы повышения достоверности маркетинговых данных // Практический маркетинг. – 2007. – № 1 – С. 20–30.

2. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследованиях // Вестник ТГЭУ. – 2006. – № 2 – С. 91–103.

3. Адасовский Б.И. Метод вычисления эталонов классов распознавания // Автоматика. 1981. – №6. – с. 3–7.
4. Орлов А.И. Эконометрика: Учеб. Пособ. для вузов / А.И.Орлов. – М.: Издательство “Экзамен”, 2002.- 576 с.
5. Мартышенко С.Н., Мартышенко Н.С., Кустов Д.А. Моделирование многомерных данных // Техника и технология. – 2007. – № 2 – С. 47–52.