Известия высших учебных заведений. Поволжский регион. – 2006. – №6(27). С. 170–177.

ЦЕНЗУРИРОВАНИЕ ПРИ ОБРАБОТКЕ АНКЕТНЫХ ДАННЫХ

С.Н. Мартышенко, Н.С., Мартышенко, Д.А. Кустов

Для повышения уровня обоснованности управленческих решений на всех уровнях управления требуется качественная и достоверная информация. Одним из основных источников первичных данных в экономических и социологических исследованиях служат данные анкетных опросов.

В последние годы в нашей стране стала бурно развиваться такая область экономики как маркетинг. Принятие решений на основе маркетинговых исследований переходит с уровня научного исследования на уровень практики все большего и большего количества предприятий. В практике маркетинговых исследований ощущается острая потребность в современных методиках и инструментальных средствах обработки анкетных данных.

Присутствующие на рынке зарубежные пакеты по обработке статистических данных являются далеко не идеальным средством анализа информации. Они больше приспособлены для применения классических статистических методов анализа к данным числовой природы, к тому же требуют чистых данных.

Такие данные больше существуют в теории, либо формируются в процессе регистрации учетной информации. Регистрировать можно, например, объемы продаж, данные о клиентах. Данные же, полученные в результате анкетных опросов, имеют принципиальные отличия. Во-первых, они содержат пропуски, во-вторых, содержат ошибки и выбросы, обусловленные множеством причин, то есть имеют сложную структуру ошибки, в-третьих, признаки, описывающие наблюдение анкеты различны по своей природе и, как правило, не относятся к признакам числовой природы или носят смешанный характер.

Современный подход к анализу анкетных данных предполагает использование многомерного статистического анализа (МСА).

Данные анкетных опросов можно рассматривать, как наблюдения многомерной случайной величины. Ответы на вопросы могут быть представлены в виде

некоторой таблицы данных, в которой строки представляют собой объекты (анкеты), а столбцы значения признаков (ответы на вопросы).

Таким образом, сама структура данных опросов содержит предпосылки применения многомерных статистических методов (МСМ). Но применение того или иного метода требует соблюдения ряда условий или требований к информации. И в первую очередь, согласованность по типам данных (непрерывные, ранговые). Одни методы предназначены для работы с одним типом данных, другие с другим. Как правило, в анкетах содержатся вопросы, порождающие множество типов данных. В зависимости от наличия групп однотипных данных исследователь выбирает и методы их обработки.

Все статистические методы предполагают наличие некоторой идеализированной выборки. Поэтому прежде чем использовать тот или иной метод анализа необходимо произвести определенную подготовительную работу – **подготовительный этап формирования данных**. Кроме того, применение метода обработки предполагает использование того или иного программного средства, которое также выдвигает свои требования к структуре и компьютерному представлению данных.

В настоящей работе мы рассмотрим только одну из задач предварительного этапа формирования данных. Она связана с первой проблемой, с которой сталкивается исследователь в реальной ситуации перед выбором методов и средств многомерного анализа - это отсутствие ответов на некоторые вопросы. Причем в одних анкетах могут отсутствовать ответы на одни вопросы в других на другие. Это приводит к отсутствию данных в таблице данных, то есть, некоторые ячейки таблицы остаются незаполненными. Игнорировать проблему отсутствия данных в реальных исследованиях невозможно. Тем более, отсутствие данных это тоже информация. Эта информация требует своего логического объяснения и разработки специальных методов анализа.

Анализ на отсутствие данных и принятие решений по обработке такой ситуации можно выстроить на основе следующих рассуждений. Зададимся вопросом, как поступить с наблюдениями, содержащими признак «отсутствие данных».

Можно вообще исключить такие анкеты из дальнейшего рассмотрения. Однако, при большом количестве вопросов в анкете, таких анкет может оказаться значительный процент. Если отсутствие данных носит не случайный характер, то при отбрасывании части анкет можно сильно исказить выборку и в конечном итоге, прийти к ошибочным выводам. Другая крайняя ситуация оставить все как есть и пытаться применять статистические методы к имеющимся данным.

При отбрасывании из таблицы данных строк с пропусками необходимо принять во внимание то, что по данным анкет решается множество задач с использованием различных статистических методов. Каждая такая задача использует далеко не все признаки. Поэтому при решении отдельной задачи, по причине отсутствия данных необходимо исключить из таблицы данных не очень большое количество строк. Тем более, нет необходимости отбрасывать все строки, в которых встречаются пропуски хотя бы по одному признаку.

Возможен и третий подход, который представляется более предпочтительным по сравнению с двумя крайними решениями. Он состоит в том, что выделяется группа анкет, которые содержат наибольшее количество пропусков. Эти анкеты подвергаются углубленному содержательному анализу, после которого исследователь принимает решение исключить такие анкеты из рассмотрения или оставить. При больших выборках, исчисляемых тысячами наблюдений, включающими до ста и более признаков подвергнуть глубокому анализу каждое наблюдение (анкету) не представляется возможным. Такие алгоритмы работают по принципу фильтров [1].

Фильтр выделяет критические анкеты, а окончательное решение остается за исследователем. Фильтр облегчает работу исследователя, сосредотачивая его внимание на критических ситуациях, то есть автоматизирует его работу.

Исследователь не в состоянии подвергнуть глубокому содержательному анализу очень большое количество анкет, а тем более произвести их сравнение без каких либо формализованных критериев. Теперь можно определить такие критерии. Простое суммирование количества отсутствующих данных по всем

признакам одного наблюдения (анкеты) не лучший критерий, поскольку признаки неравнозначны с точки зрения восприятия вопросов анкеты респондентами.

Для подтверждения этого утверждения рассмотрим два вопроса анкеты. Предположим, что на один вопрос не дали ответ 50% респондентов и отсутствие данных в этом случае почти норма. На другой вопрос не дали ответ 5% респондентов, и отсутствие данных в этом случае, требует анализа возможных причин такой ситуации. Поэтому в критерии целесообразно учесть вес вопроса с точки зрения восприятия вопроса всеми респондентами.

Дадим формализованное представление критерия. Обозначим строку таблицы данных, как:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{im}), \tag{1}$$

где i – номер объекта, i=1,2,…,n;

n – количество анкет;

j – номер вопроса (признака), j=1,2,…,m;

m – количество признаков.

Для упрощения записи введем переменную ν_{ij}:

$$\nu_{ij} = \begin{cases} 0 - \text{если } x_{ij} \text{ - есть данные} \\ 1 - \text{если } x_{ij} \text{ - нет данных} \end{cases} \tag{2}$$

Тогда критерий отбора подозрительных анкет можно записать в виде:

$$\varphi_i = \sum_{j=1}^{m} Q_j \nu_{ij}\ , \tag{3}$$

где Q_j - весовой коэффициент вопроса, который рассчитывается путем нормировки коэффициентов q_j:

$$q_j = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \nu_{ij}}{\sum_{i=1}^{n} \nu_{ij}} \tag{4}$$

То есть:

$$Q_j = \frac{q_j}{\sum_{j=1}^{n} q_j} \qquad (5)$$

Весовой коэффициент Q_j вводится для ранжирования признаков. Если признак содержит множество незаполненных позиций (пропуски данных), то отсутствие данных признака в отдельном наблюдении событие не столь уж исключительное и имеет небольшой вес. После расчета φ_i (i=1, 2, …, n) выборка может быть упорядочена по убыванию показателя φ_i. На первых позициях окажутся наблюдения, внушающие наибольшее беспокойство из-за наличия отсутствующих данных. Анкеты, соответствующие таким наблюдениям, должны быть подвергнуты углубленному содержательному анализу.

Такой подход позволяет найти и отбросить анкеты, резко отличающиеся от всех остальных. Если в процессе анализа возникает подозрение, что пропуски вызваны негативным отношением респондентов к опросу, то такие анкеты должны быть исключены. Степень уверенности в необходимости исключения таких анкет повышается, если в поле зрения исследователя попадает серия анкет, принадлежащих одному пакету, предоставленному одним интервьюером. Причину отсутствия данных в этом случае нельзя признать случайной. Практика показывает, что и другие ответы в таких анкетах не отличаются высокой достоверностью. Поэтому такие анкеты лучше вообще исключить, как грубые выбросы. На практике, количество бракуемых с помощью этого критерия анкет не превосходит 2-3 % и при выборках, исчисляемых тысячами анкет, такая потеря не ухудшает картины, а качество информации и возможности обработки многомерных данных повышаются.

Рассмотренный фильтр является далеко не единственным инструментом анализа качества данных. Для получения более достоверного результата, данные целесообразно подвергнуть анализу с помощью нескольких фильтров. Чтобы отличать одни фильтры от других будем давать им специальные названия. Рассмотренный фильтр назовем – **“Фильтр отсутствия данных” (ФОД).**

Проблема отсутствия данных стоит не только для отдельного наблюдения, но и для отдельного интервьюера. Может оказаться, что из-за некачественной ра-

боты одного из интервьюеров весь пакет анкет такого интервьюера резко отличается от пакетов, собранных другими интервьюерами. При больших объемах выборки с привлечением десятков интервьюеров возможно лучше вообще отказаться от всего пакета, представленного недобросовестным работником. Для того чтобы ввести оценку интервьюера по отсутствию данных, необходимо ввести новые обозначения данных:

$$x_{i_r} = (x_{i_r 1}, x_{i_r 2}, \ldots, x_{i_r j}, \ldots, x_{i_r m}) \tag{6}$$

где j=1,2,…,m – номер признака; m - количество признаков;

r=1, 2, …, k – номер интервьюера, k – количество интервьюеров;

i_r=1, 2, …, n_r – номер анкеты в пакете одного интервьюера;

n_r – объем пакета анкет интервьюера с номером r;.

Тогда, объем выборки, включающей все анкеты будет:

$$n_0 = \sum_{r=1}^{k} n_r \ . \tag{7}$$

Задача состоит в том, чтобы из k пакетов анкет выделить пакет, который имеет наибольшие отличия от остальных пакетов.

Обозначим оценку пакета r-го интервьюера как:

$$w_r = \sum_{i_r=1}^{n_r} \varphi_{i_r} / n_r \tag{8}$$

С содержательной точки зрения, оценка w_r представляет собой осреднение значения оценки φ_i по данным пакета, предоставленного r- ым интервьюером.

Однако, этой величины недостаточно для ранжирования интервьюеров. Ее необходимо сопоставить с величиной средней оценки, рассчитанной по всем остальным интервьюерам:

$$w_{-r} = \frac{\sum_{r=1}^{k} \sum_{i_r=1}^{n_r} \varphi_{i_r} - \sum_{i_r=1}^{n_r} \varphi_{i_r}}{\sum_{r=1}^{k} n_r - n_r} \tag{9}$$

Тогда, в качестве подозрительного можно назвать интервьюера, для которого критерий:

$$V_r = w_{-r} - w_r \qquad (10)$$

принимает минимальное значение.

В результате дополнительного анализа выделенных с помощью критерия пакетов, некоторые пакеты могут быть исключены из рассмотрения, что не только повысит достоверность статистического вывода, но сделает данные более пригодными для применения методов МСА.

Рассмотренный фильтр назовем **”Фильтр отсутствия данных групповой” (ФОДГ)**.

Во многих случаях отсутствующие данные удается восстановить с известной степенью точности. В ряде статистических пакетов содержатся средства восстановления пропусков в отдельно взятом признаке. Такие процедуры применимы только к признакам числовой природы. Кроме того, одномерные методы восстановления данных часто могут приводить к абсурдным результатам с точки зрения логической связи с другими признаками.

Нами был разработан ряд алгоритмов восстановления пропущенных данных по многомерной выборке. Эти алгоритмы можно разделить на две группы - логические и статистические. В данной работе рассмотрим основную идею статистических методов.

Основная идея статистического восстановления состоит в применении методов многомерной статистической классификации и распознавания образов. В силу того, что анкетные данные часто содержат признаки нечисловой природы или смешанные данные, для решения задачи восстановления пропусков применимы далеко не всякие алгоритмы распознавания.

Для работы с признаками, измеренными в различных шкалах, мы использовали метод распознавания образов, разработанный Б.И. Адасовским в 80 –ые годы прошлого века. Этот метод получил название – метод интегральной диагностики многомерных систем. Основные теоретические положения, заложенные в основу метода интегральной диагностики многомерных систем, изложены в работах [1; 2]. Идея метода заключается в вычислении эталонов для классов заданного разбиения обучающей выборки в спрямляющем булевом пространстве. Однако, при

всей универсальности подхода этот метод нашел очень ограниченное применение в практике. Он был использован в основном для разработки технических систем.

Рассмотрим принцип восстановления данных на примере одного из признаков, например, признака с номером *j*. Предполагается, что признак *j* имеет определенное количество возможных значений L_j (дискрет). Процедура восстановления включает два основных этапа – обучение и распознавание. На первом этапе в качестве обучающей выборки используются многомерные данные, не содержащие пропусков. Номера дискрет можно положить в основу разбиения выборки данных на классы. Классифицированная выборка служит в качестве обучающей выборки, по которой строится эталон классов. Признак *j* не участвует в построении эталона. Восстановление данных производится при распознавании контрольной выборки, составленной из наблюдений с пропусками. Восстанавливая номер класса для элементов контрольной выборки, мы тем самым определяем номер дискрета или некоторое отсутствующее значение.

Однако, прежде чем использовать процедуру распознавания для восстановления данных, необходимо оценить уровень погрешности восстановления пропусков. Для оценки матрицы ошибок мы применяем процедуру скользящего экзамена к данным обучающей выборки. Если процент ошибок не очень высок, то алгоритм можно использовать для восстановления значений признака.

Исследование эффективности работы программных модулей, реализующих статистические фильтры, продемонстрируем на примере работы с данными анкетного опроса по изучению пляжно-оздоровительного отдыха в Приморском крае. Количество респондентов участвовавших в опросе – 3361 человек. Вопросы анкеты описываются 72 основными признаками. Текст анкеты и опрос был произведен старшим преподавателем кафедры маркетинга ВГУЭС Н.С. Мартышенко.

Результатом работы фильтра является список значений критерия ФОД. После расчета критерия анкеты упорядочиваются по убыванию его значений. Для визуализации результатов строим график по первым 30 – 40 наибольшим значениям критерия (рис.1). Из графика видно, что первые 11 точек отличаются от всех остальных по значению критерия ФОД.

Проанализируем причины возникновения выбросов. Для этого воспользуемся дополнительной информацией, предоставляемой программным модулем, реализующим ФОД.

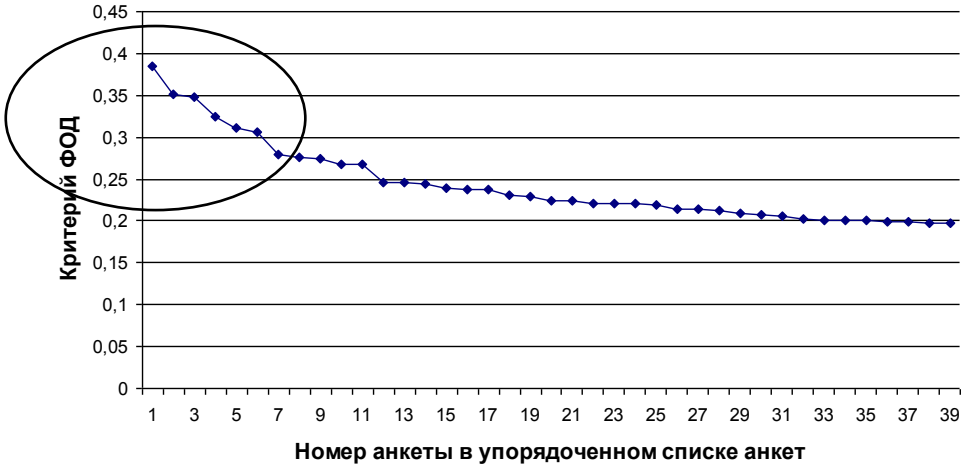

Рис. 1. Значения критерия ФОД для верхней части упорядоченного списка анкет

В данном случае максимальный вклад в критерий ФОД внес признак “До пляжа добираюсь (используемое транспортное средство)”, из 3361 анкеты по этому признаку отсутствуют данные только в 17 анкетах (0,5%), поэтому ошибка в этом признаке получила наибольший вес - 0,15416.

Но с содержательной точки зрения, этот признак легко восстановить по остальным данным с помощью логических процедур.

После восстановления значений признака “До пляжа добираюсь” повторно запускаем программный модуль ФОД. Результаты расчетов приведены на рис. 2.

На графике (рис. 2) первые 7 точек отличаются от остальных. В данном случае максимальный вклад в критерий ФОД внес признак “Затраты времени на дорогу”, из 3361 анкеты по этому признаку отсутствуют данные только в 31 анкете (0,92%), поэтому ошибка в этом признаке получила наибольший вес - 0,099945. Пропущенные данные также можно восстановить с допустимым уровнем погрешности.

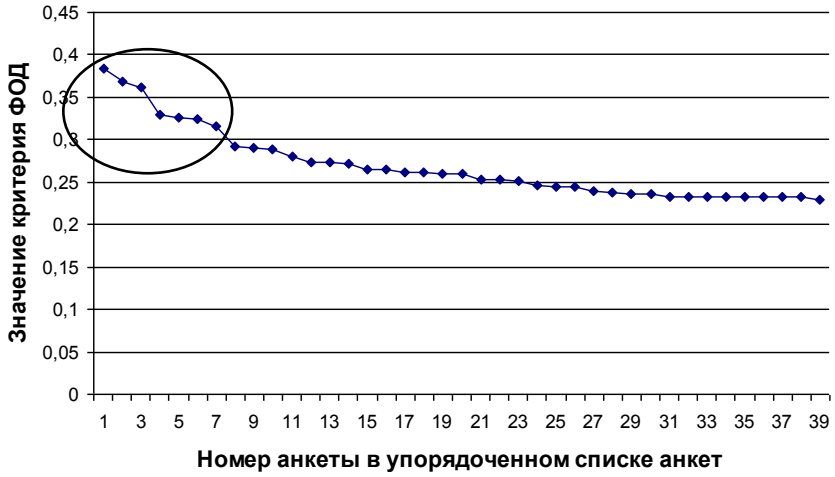

Рис. 2. Значения критерия ФОД для верхней части упорядоченного списка анкет при повторном запуске программы

Теперь рассмотрим работу фильтра отсутствия данных при анализе пакетов анкет. Полученные результаты отсортируем по возрастанию. В результате чего получим упорядоченный список пакетов. Построим график по первым 30 - 40 наименьшим значениям критерия ФОДГ (рис. 3).

Рис. 3. Значения критерия ФОДГ для верхней части упорядоченного списка анкет

Из графика видно, что первые 3 точки достаточно сильно отличаются от остальных. Эти пакеты содержат 37 анкет (0,11%), поэтому мы исключим их из таблицы данных и повторим запуск программного модуля. Оставшиеся пакеты анкет дают близкие значения группового критерия.

Программные модули, позволяющие обрабатывать многомерные данные с пропусками, входят в состав разработанного авторами специализированного па-

кета анализа анкетных данных. Данный пакет разрабатывался, как приложение к EXCEL. Возможность использования программных модулей в среде EXCEL делают его доступным для самого широкого круга практиков.

Литература

1. Адасовский Б.И. Метод вычисления эталонов классов распознавания // Автоматика. 1981. – №6. – с. 3–7.

2. Адасовский Б.И. О мере близости классов распознавания // Кибернетика – 1986. – №4. – с. 116-117.

3. Мартышенко Н.С. Методическое обеспечение анализа поведения потребителей на региональном туристском рынке // Вестник Тихоокеанского государственного экономического университета. – 2005. - №4. С. 19-31.

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

5. Мартышенко С.Н., Д.А. Кустов Применение методов многомерной классификации для анализа данных анкетных опросов // Интеллектуальный потенциал вузов – на развитие Дальневосточного региона России: материалы VIII Международной конференции аспирантов и молодых исследователей, Владивосток, 2006.

6. Мартышенко Н.С., Д.А. Кустов Методика повышения достоверности анкетных данных // Интеллектуальный потенциал вузов – на развитие Дальневосточного региона России: материалы VIII Международной конференции аспирантов и молодых исследователей, Владивосток, 2006.