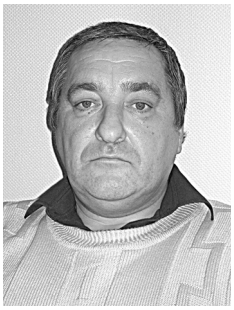
Сергей МАРТЫШЕНКО,
Наталья МАРТЫШЕНКО,
Дмитрий КУСТОВ

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ МАРКЕТИНГОВЫХ ДАННЫХ



Мартышенко Сергей Николаевич – к. т. н., проф. кафедры математики и моделирования Владивостокского государственного университета экономики и сервиса.

Мартышенко Наталья Степановна – к. э. н., ст. преподаватель кафедры маркетинга и коммерции Владивостокского государственного университета экономики и сервиса.

Кустов Дмитрий Александрович – ст. преподаватель кафедры информатики, инженерной и компьютерной графики Владивостокского государственного университета экономики и сервиса.

Статья принята к печати 8 ноября 2006 года.

Основным источником первичных данных в маркетинговых исследованиях служат данные анкетных опросов. Однако многие исследования, основанные на анализе анкетных данных, завершаются уже на этапе предварительного анализа, который зачастую заключается в анализе частотных рядов для одномерных признаков. То есть огромное количество полезной информации, содержащейся в данных, остается неиспользованной. Современный подход к анализу анкетных данных предполагает использование многомерного статистического анализа.

Данные анкетных опросов можно рассматривать как наблюдения многомерной случайной величины (т*абл.*). 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 анкетных опросов содержит предпосылки применения многомерных статистических методов. Но применение того или иного метода требует соблюдения ряда условий или требований к данным. 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. Многообразие типов данных обуслов-

лено разнообразием шкал, используемых в анкетах для измерения характеристик объектов исследования. Наличие разнообразных шкал вызвано не прихотью исследователей, а их стремлением получить от респондентов более достоверную информацию. Поскольку не респондент, а исследователь заинтересован в получении информации, ему приходится подстраиваться под респондента, предоставляя вопросы в удобной для респондента форме.

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 данных предполагает использование того или иного программного средства, которое также выдвигает свои требования к структуре и компьютерному представлению данных.

Первая проблема, с которой сталкивается исследователь в реальной ситуации перед выбором методов и средств многомерного анализа данных анкет — 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. Такой анализ имеет особую ценность и для организации последующих исследований. Для определения анализа отсутствия данных введем специальный термин — **первая проблема подготовительного этапа формирования данных**.

Таблица

Табличная форма представления данных анкетного опроса

Номер	Признак 1	Признак 2	....	Признак *j*	....	Признак *m*
1	x_{11}	x_{12}	....	x_{1j}	....	x_{1m}
2	x_{21}	x_{21}	....	x_{2j}	....	x_{2m}
....	....	....	....	....	....	....
i	x_{i1}	x_{i2}	....	x_{ij}	....	x_{im}
....	....	....	....	....	....	....
N	x_{n1}	x_{n2}	....	x_{nj}	....	x_{nm}

Однако прежде чем перейти к рассмотрению идей анализа отсутствия данных, необходимо установить некоторые правила описания и компьютерного представления отсутствия данных. Необходимость единообразия описания ситуации отсутствия данных обусловлена еще и требованиями системного подхода к разработке компьютерных технологий обработки данных, которые строятся с учетом некоторых общих свойств данных и специфики решаемых задач. Компьютерные технологии воплощаются в виде комплекса проблемно-ориентированных программных средств позволяющих наиболее эффективно решать задачи, объединенные единой предметной областью и общей системой данных.

К сожалению, единого компьютерного представления ситуации отсутствия данных для различных типов данных дать не удается. Так, например, пустая ячейка при обработке числовых данных большинством универсальных программных комплексов воспринимается как нулевое значение. Но при ответах на вопросы анкет нулевое значение может быть одним из возможных вариантов ответа, то есть вполне информативным. Неоднозначность ситуации восприятия отсутствия данных возникает из-за того, что классификация компьютерных типов данных (текстовые, с плавающей точкой, целые даты и др.) не совпадает с содержательной классификацией данных порождаемых анкетными опросами или измерительными шкалами (номинальная, порядковая, интервальная, относительная и др.). Поэтому если не удается прийти к единообразному представлению ситуации отсутствия данных для различных признаков, то

необходимо стремиться хотя бы минимизировать количество возможных вариантов.

В большинстве случаев данные анкетных опросов можно представить в компьютере текстовыми и числовыми полями. Одним из вариантов описания может быть вариант, использованный нами при разработке специализированного комплекса программных средств анализа анкетных данных[1]. Отсутствие данных в текстовом поле обозначается «н/д», а в числовом «-1». Кодирование отсутствия данных в числовом поле как «-1» вполне оправданно для анкетных опросов в социологических и маркетинговых исследованиях, поскольку значение признака «-1» представить достаточно трудно.

Данные могут отсутствовать и в вопросе, требующем указать дату. Но такие вопросы в анкетах встречаются крайне редко. В компьютерном представлении такое поле может быть закодировано, например, датой, выходящей за пределы возможных значений.

Существует группа вопросов и, соответственно, данных, для которых вообще невозможно однозначно определить признаки отсутствия данных. Например, вопрос «Сколько раз в год вы посещаете сауну?». Если респондент ничего не указал в графе ответа, то мы не можем предполагать, что он отказался от ответа, а вынуждены предполагать, что сауну он не посещает, то есть ответ «0», а это уже не отсутствие данных. Для таких вопросов признака отсутствия данных вообще нет. Другими словами, считается, что данные всегда есть. Однако в зависимости от цели, стоящей перед исследователем, он может принять и другое решение.

Существует группа вопросов, для которых отсутствие данных можно определить только в результате сопоставления ответа с ответом на другой вопрос. Такие вопросы можно назвать зависимыми вопросами от некоторого главного.

Например, если в анкете присутствует вопрос о занятости (работе), в котором требуется выбрать ответ из двух альтернатив: «да», «нет». Далее в анкете идут вопросы о транспорте, который используется для проезда на работу, условиях работы и т. д. Отсутствие данных при ответе на зависимый вопрос может быть вполне естественным, если ответ на главный вопрос «нет».

С таким вопросом необходимо различать две ситуации: первая ситуация возникает, когда респондент на главный вопрос дал ответ «да», а на зависимый вопрос не отвечал; вторая ситуация возникает, когда на главный вопрос получен ответ «нет» и на зависимый вопрос ответ тоже не указан.

В первом случае ситуация отсутствия данных обрабатывается точно так же, как для обычных текстовых и числовых данных, то есть кодируется либо «н/д» либо «-1». Во втором случае для всех зависимых от главного текстовых вопросов устанавливается некоторый общий ответ. Для нашего примера это может быть, например, текст: «не работаю». Для числовых вопросов в большинстве случаев ситуация легко описывается значением ноль — «0». Действительно, для нашего примера это будет: продолжительность рабочего дня ноль часов, заработная плата ноль рублей и т. д.

В анкетах могут встретиться такие зависимые вопросы, которые не имеют главного вопроса, то есть зависимы между собой. Логическое предположение об отсутствии данных в этом случае можно сделать только по группе ответов на зависимые вопросы. В таких ситуациях полной определенности часто добиться не удается.

Например, в вопросе просят указать сезон отпуска за последние три года в трех позициях. Если респондент указал сезон отпуска только в одной или двух позициях, а в остальных не указал, то логично предположить, что отпуска в соответствующий год не было. Если респондент не указал сезон отпуска за все три года, то такую ситуацию логичнее описать, как отсутствие данных во всех трех позициях, поскольку отсутствие отпуска подряд три года маловероятно.

[1] *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.

В любом случае, некоторая неопределенность в интерпретации ситуации отсутствия данных остается, поскольку исследователь должен самостоятельно определить ситуацию, основываясь на собственных оценках вероятности событий, отдавая предпочтение более вероятным событиям по отношению к менее вероятным. Однако, выработав определенную логику, исследователь обязан применять ее для всех объектов.

Во избежание ошибок ситуации с зависимыми вопросами, как с главным вопросом, так и без него, целесообразно протестировать с помощью специальных логических фильтров (такие фильтры имеются в нашем комплексе программных средств). Иначе одна и та же ситуация может быть определена при вводе данных в компьютер различными способами.

При массовом анкетировании и большом количестве вопросов анкеты ожидать, что отсутствие данных не возникнет, не приходится. Решение вопроса описания ситуаций отсутствия данных при ответах на вопросы анкеты должно быть произведено еще на этапе разработки текста анкеты, а затем конкретизировано при разработке программных средств, используемых для ввода данных в компьютер. Иначе будет невозможно разработать специализированные программные средства, предназначенных для обработки таких ситуаций.

Теперь, после формализации ситуации отсутствия данных в анкете, можно перейти к обсуждению вопросов анализа и компьютерной обработки таких данных.

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 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 отсутствием данных необходимо принять во внимание то, что по анкетам решается множество задач с использованием различных статистических методов. Каждая такая задача использует далеко не все признаки. Поэтому при решении отдельной задачи необходимо исключить из таблицы данных по причине отсутствия данных не очень большое количество строк. Тем более, нет необходимости отбрасывать все строки, в которых встречается отсутствие данных 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 значений отсутствия данных. Эти анкеты подвергаются углубленному содержательному анализу, после которого исследователь принимает решение, исключить такие анкеты из рассмотрения или оставить. Этот алгоритм работает по принципу фильтра. В отдельных случаях исследователь может с большой степенью достоверности восстановить отсутствующие данные на основании ответов на другие вопросы анкеты.

Фильтр выделяет критические анкеты, а окончательное решение остается за исследователем. Фильтр облегчает работу исследователя, сосредоточивая 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. Идея ранжирования анкетных данных с помощью формализованного критерия была выдвинута в работе[2]. Одна-

[2] *Мартышенко Н.С.* Методическое обеспечение анализа поведения потребителей на региональном туристском рынке/ Н.С. Мартышенко // Вестник ТГЭУ. 2005. № 4. С. 19–31.

ко в этой работе предложено всего несколько простейших фильтров, которые не были воплощены в виде конечных программных продуктов, пригодных для широкого практического использования. В настоящее время нами разработана и реализована целая система фильтров, а также произведено исследование возможностей различных фильтров на различных анкетных данных.

Теперь можно определить критерии ранжирования анкет по признаку отсутствия данных. 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

$x_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im})$, (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 v_{ij}:

$$v_{ij} = \begin{cases} 0 - \text{если } x_{ij} \text{ - есть данные} \\ 1 - \text{если } x_{ij} \text{ - нет данных} \end{cases} \quad (2)$$

Тогда критерий отбора подозрительных анкет можно записать в виде:

$$\varphi_i = \sum_{j=1}^{m} Q_j v_{ij}, \quad (3)$$

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

$$q_j = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} v_{ij}}{\sum_{i=1}^{n} v_{ij}}. \quad (4)$$

То есть $$Q_j = \frac{q_j}{\sum_{j=1}^{n} q_j}. \quad (5)$$

Весовой коэффициент Q_j вводится для ранжирования признаков. Если признак содержит множество незаполненных позиций (отсутствие данных), то отсутствие данных признака в отдельном наблюдении событие не столь уж исключительное и имеет небольшой вес и наоборот. После расчета (i=1, 2, ..., n) выборка может быть упорядочена по убыванию показателя φ_i. На первых позициях окажутся наблюдения, внушающие наибольшее беспокойство наличием 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 отсутствие данных вызвано негативным отношением респондентов к опросу, то такие анкеты должны быть исключены. Степень уверенности в необходимости исключения данных анкеты 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 данных и возможности их обработки с помощью многомерных методов облегчаются.

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

$$x_{i_r} = (x_{i_r 1}, x_{i_r 2}, ..., x_{i_r j}, ..., x_{i_r m}), \quad (6)$$

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:

$$n_0 = \sum_{r=1}^{k} n_r \quad . \quad (7)$$

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:

$$w_r = \sum_{i_r=1}^{n_r} \varphi_{i_r} / n_r \, . \quad (8)$$

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

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

$$w_{-r} = \frac{\sum_{r=1}^{k} \sum_{i_r=1}^{n_r} \varphi_{i_r} - \sum_{i_r=1}^{n_r} \varphi_{i_r}}{n_0 - n_r} \, . \quad (9)$$

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

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

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 многомерного анализа.

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

Возможности фильтра **ФОДГ** продемонстрируем на примере обработки анкет по изучению пляжно-оздоровительного отдыха. В опросе участвовали 3 361 респондент. Вопросы анкеты описываются 72 основными признаками. Опрос был произведен Н.С. Мартышенко. Упорядоченные значения критерия **ФОДГ** представлены на *рисунке 1*.

Из графика (рис. 1) видно, что первые 3 точки достаточно сильно отличаются от остальных. Эти пакеты содержат 37 анкет (0,11%), поэтому мы исключили их из таблиц данных. После этого мы повторили запуск программы и получили плавную кривую значения критерия.

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

Теперь исследователь должен подвергнуть данные критическому анализу на наличие больших выбросов, которые могут возникнуть из-за грубых ошибок. Эту проблему можно определить как **вторую проблему подготовительного этапа** анализа анкетных данных.

Как и в случае анализа таблицы данных на отсутствие данных, анализ с целью выявления грубых выбросов — данных, резко отличающихся от основной массы, предлагается производить не по отдельным значениям признака, а по многомерным векторам признаков. При выявлении грубых выбросов возникают вопросы: являются ли отклоняющиеся данные действительно ошибками или это реальные значения и как получить адекватные оценки параметров изучаемой совокупности? Решением таких вопросов занимается специальный раздел статистики — робастное оценивание.

В данной работе анализ выбросов многомерных данных предлагается производить по единому подходу. Суть его состоит в разработке многомерных фильтров, с помощью которых выделяется заданное число подозрительных объектов, которые затем подвергаются углубленному содержательному анализу. То есть принятие окончательного решения о том, как поступить с выбросом, остается за исследователем.

Фильтры для данных, измеренных в различных шкалах, строятся по-разному. Другими словами, для анализа используются группы признаков, измеренные в одной шкале.

Рассмотрим фильтр, предназначенный для работы с признаками, измеренными в шкале отношений. Формальное описание фильтра дадим не для отдельного объекта (анкеты), а для пакета данных интервьюера. Фильтрация отдельных наблюдений является частным случаем фильтрации пакетов при количестве наблюдений в пакете $n_r = 1$ и $k = n_0$.

Запишем одно наблюдение из r-го пакета набором m значений:

$$x_{ir} = (x_{ir1}, x_{ir2}, ..., x_{irj}, ..., x_{irm}), \quad (11)$$

считая, что все m признаков измерены в шкале отношений. Тогда задача состоит в том, чтобы из k пакетов анкет выделить пакет, который имеет наибольшие отличия от остальных пакетов.

Для этого последовательно для каждого пакета r ($r = 1, 2, ..., k$) повторим следующую процедуру: рассчитаем средние значения m признаков по выборке за исключением пакета с номером r

$$\overline{X}^{-r} = (\overline{x}_1^{-r}, \overline{x}_2^{-r}, ..., \overline{x}_j^{-r}, ..., \overline{x}_m^{-r}), \quad (12)$$

и средние значения признаков по пакету с номером r

$$\overline{X}^{r} = (\overline{x}_1^{r}, \overline{x}_2^{r}, ..., \overline{x}_j^{r}, ..., \overline{x}_m^{r}). \quad (13)$$

Вычисляем поэлементные модули разностей двух векторов средних:

$$\lambda_r = \left| \overline{X}^{-r} - \overline{X}^{r} \right|. \quad (14)$$

Процедура, основанная на последовательном изъятии и восстановлении части выборки, называется **скользящим экзаменом**. Объединяем все отклонения λ_r в одну матрицу λ размерности k x m. На основании матрицы λ рассчитаем матрицу M той же размерности. Вычисления производятся по схеме: определяется максимум в каждом столбце матрицы λ, затем элементу матрицы M, соответствующему максимуму, присваивается значение единицы, всем остальным элементам матрицы M присваивается значение ноль. В результате построчного суммирования элементов матрицы M получим вектор оценок для каждого интервьюера:

$$\mu^{M} = (\mu_1^{M}, \mu_2^{M}, ..., \mu_r^{M}, ..., \mu_k^{M}). \quad (15)$$

Интервьюер с наибольшим значением μ^{M} будет иметь максимальный штраф, и поэтому его данные могут быть поставлены под сомнение. Теперь исследователь может сосредоточить свое внимание на отдельном пакете первичных данных, подвергнуть их дополнительному содержательному анализу, в результате которого исследователь определяет, является отклонение допустимым или нет. Это достаточно грубый фильтр. Он основан на предположении о том, что если пакет анкет содержит недостоверную информацию, то большие отклонения от средних значений будут не только по одному признаку, но и по другим. Данную процедуру можно применить и в случае, когда пакеты содержат по одному наблюдению. В этом случае идентификатором интервьюера может выступить номер наблюдения в выборке.

В программной реализации данный фильтр допускает применение двух вариантов весовых коэффициентов признаков. В первом случае учитывается наличие нескольких значений признака, соответствующих максимальному значению. В этом случае весовые коэффициенты q_j рассчитываются по формуле:

$$q_j = \frac{\sum_{j=1}^{m} \gamma_j}{\gamma_j} \quad , \qquad (16)$$

где γ_j – количество значений, равных максимальному значению для признака с номером j в исходной матрице данных.

То есть веса и для группового и для обычного фильтра рассчитываются по одной формуле. Тогда коэффициенты Q_j получаются путем нормировки коэффициентов q_j:

$$Q_j = \frac{q_j}{\sum_{j=1}^{n} q_j} \qquad (17)$$

Вектор оценок интервьюеров с учетом весов, будет равен:

$$\mu^{M1} = (\mu_1^{M1}, \mu_2^{M1}, ..., \mu_r^{M1}, ..., \mu_k^{M1}) . \qquad (18)$$

Второй вариант расчета весовых коэффициентов предполагает предварительную модульную нормализацию признаков. То есть исходные значения признаков x_{ij} преобразуются к виду:

$$x'_{ij} = \frac{\left| x_{ij} - \overline{X}_j \right|}{\sqrt{S_j^{\,2}}} \quad , \qquad (19)$$

где $\overline{X}_j$ – среднее значение признака с номером j;

S^2_j – оценка дисперсии признака с номером j.

Тогда графическую интерпретацию вводимого коэффициента на примере двух признаков можно изобразить в виде рисунка (*рис.2*). Второй вариант на реальных данных показал более хорошие результаты

Рассмотренному фильтру дадим следующее название – **«Фильтр экстремальных непрерывных значений» (ФЭНЗ)**. Если рассматриваются пакеты анкет, то название дополняется словом «групповой», то есть сокращенное название будет (**ФЭНЗГ**).

Рассмотренный фильтр больше пригоден все-таки для обработки пакетов на начальной стадии анализа выбросов. Он вообще может не показать результатов, если каждый объект содержит не более одного выброса. Более тонкий фильтр можно создать на основе ранжирования отклонений от среднего значения признака.

Матрица λ в этом фильтре рассчитывается, как и в фильтре экстремальных непрерывных значений. Затем матрица λ преобразуется в матрицу рангов λ_p. То есть минимальному значению в столбце присваивается значение 1. Следующему значению – 2 и так далее. Наибольшему значению в столбце присваивается значение k. Повторяющимся значениям сопоставляется ранг, равный среднему значению рангов повторяющихся значений.

То есть сумма рангов каждого столбца матрицы λ_p будет равна сумме ряда натуральных чисел от 1 до k. В результате построчного суммирования элементов матрицы λ_p получим оценки интервьюеров

$$\mu^p = (\mu^p_1, \mu^p_2, ..., \mu^p_k). \qquad (20)$$

Этот фильтр дает неплохие результаты, когда нет очень тесной связи между всеми признаками. При большом количестве признаков такой зависимости, как правило, не обнаруживается.

Определим этот фильтр, как **«Фильтр ранжирования непрерывных значений» (ФРНЗ),** и при рассмотрении пакетов – (**ФРНЗГ**) соответственно.

Кроме рассмотренных фильтров, для непрерывных данных можно было бы определить се-

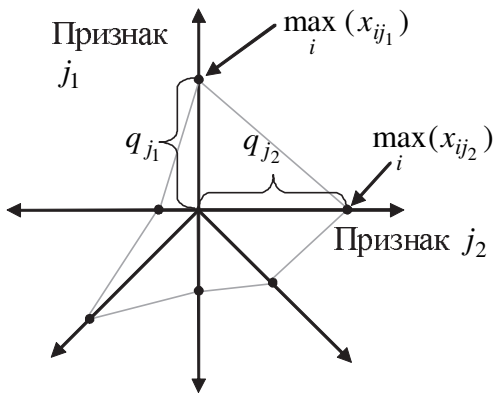


Рис. 2. Графическая интерпретация весового коэффициента

рию метрических фильтров, основанных на подсчете расстояния от объекта до центра выборки с использованием известных метрик Евклида, Хэмминга и Махаланобиса:

$$\rho_{\text{Е}} = \sqrt{\sum_{j=1}^{m}(x_{ij} - \bar{x}_j^{-i})^2}\ , \quad (21)$$

$$\rho_{\text{Х}} = \sum_{j=1}^{m}\left|(x_{ij} - \bar{x}_j^{-i})\right|, \quad (22)$$

$$\rho_{\text{М}} = (x_i - \bar{X}^{-i})'\Sigma^{-1}(x_i - \bar{X}^{-i}), \quad (23)$$

где $\rho_{\text{Е}}$, $\rho_{\text{Х}}$, $\rho_{\text{М}}$ – Евклидово, Хэммингово и Махаланобиса расстояния от отдельного объекта, заданного вектором признаков x_i, до центра выборки, рассчитанного при исключении объекта с номером *i*;

$x_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im})$ – вектор признаков *i*-го объекта, $i = 1, 2, ..., n$;

$\bar{X}^{-i} = (\bar{x}_1^{-i}, \bar{x}_2^{-i}, ..., \bar{x}_j^{-i}, ..., \bar{x}_1^{-i})$ – центр выборки, рассчитанный без учета объекта с номером *i*;

Σ^{-1} – матрица. обратная к ковариационной матрице выборки, рассчитанной без учета объекта с номером *i*.

Объекты, наиболее удаленные от центра выборки, определенного в многомерном пространстве признаков, можно было бы тоже выделить как нетипичные. Фильтры, построенные на подсчете расстояния до центра выборки, определим под общим названием **«Фильтр метрический непрерывных значений» (ФМНЗ)** и при рассмотрении пакетов – (**ФМНЗГ**) соответственно.

Для признаков, измеренных в ранговых и номинальных шкалах, можно использовать фильтр, основанный на сравнении частотных рядов признаков. К номинальным шкалам можно отнести и бинарные признаки. К номинальной шкале с помощью операции **типизации**, которая также представлена в нашем пакете, могут быть преобразованы и ответы на открытые вопросы. Практика показывает их высокую информативность.

Предположим, что всего выделено *m* признаков, измеренных в номинальной шкале. Тогда для расчета элемента вектора λ_r будем использовать формулу:

$$\lambda_{rj} = \sum_{t_j=1}^{f_j}\frac{(P^{-r}{}_{t_j j} - P^{r}{}_{t_j j})^2}{P^{-r}{}_{t_j j}}\ , \quad (24)$$

где t_j – номер интервала частотного ряда *j*-го признака $t_j = 1, 2, ..., f_j$;

f_j – количество интервалов в частотном ряду *j*-го признака;

$P^{r}t_j j$, $P^{r}t_j j$ – относительные частоты, рассчитанные по частотным рядам выборки без *r*-го пакета и только по выборке *r*-го пакета.

Поскольку количество разрядов в частотных рядах различных признаков может отличаться, то в общей оценке интервьюера признаки должны браться с весовыми коэффициентами:

$$\mu_r = \sum_{j=1}^{m}\frac{f_j}{\sum_{j=1}^{m} f_j}\lambda_{rj} \quad (25)$$

Необходимость весовых коэффициентов отпадает, если все признаки свести к бинарным шкалам. Соответственно можно использовать этот фильтр на всех признаках, если предварительно перейти от шкалы отношений к ранговой или бинарной шкале.

Определим этот фильтр, как **«Фильтр частот кодированных значений» (ФЧКЗ)**, и при рассмотрении пакетов – (**ФЧКЗГ**) соответственно. Этот фильтр, так же как и фильтр ФЭНЗ для непрерывных значений, больше предназначен для обработки пакетов из кодированных признаков.

Рассмотрим еще один фильтр для данных, измеренных в номинальной шкале относительно отдельных наблюдений. В этом случае одно наблюдение мы рассмотрим как

$x_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., xi_m)$. (26)

При этом считаем, что все m признаков представлены в номинальной шкале. В номинальной шкале каждый признак может быть представлен некоторым ограниченным списком уникальных значений

$$a_j = a_1, a_2, ..., a_{tj}, ..., a_{fj}, \quad (27)$$

где j – номер признака $j = 1, 2, ..., m$;

t_j – номер уникального значения в списке уникальных значений *j*-го признака, $t_j = 1, 2, ..., f_j$;

f_j – количество уникальных значений в списке уникальных значений *j*-го признака.

Таким образом, любой элемент x_{ij} вектора x_i принимает одно из значений множества a_j. С каждым значением a_{tj} можно сопоставить частоту встречаемости такого значения в *j*-м признаке выборки:

$\pi_j = (\pi_1, \pi_2, ..., \pi_{tj}, ..., \pi_{fj})$, (28)

где $\pi_{ti} = n_{ti}/n_0$; n_{ti} – количество значений a_{tj} в столбце j.

Произведем замену значений x_{ij} на соответствующие им значения из множества π_j. Тогда каждое наблюдение выборки будет заменено вектором

$$\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{ij}, ..., \theta_{im}), \quad (29)$$

где

$$\theta_{ij} \in \pi_j. \quad (30)$$

В общем случае сумма θ_{ij} не равна единице:

$$\Omega_i = \sum_{i=1}^{m} \theta_{ij} \neq 1 \ . \quad (31)$$

Значение (Ω_i/m) можно использовать в качестве отбора подозрительных объектов выборки. Наиболее обособленному объекту выборки должно соответствовать минимальное значение (Ω_i/m).

Определим этот фильтр, как **«Фильтр замены кодированных значений» (ФЗКЗ)** и при рассмотрении пакетов – (**ФЗКЗГ**) соответственно.

Кроме рассмотренных фильтров наш программный комплекс включает некоторый универсальный фильтр, особенность которого состоит в том, что он может работать со всеми типами признаков, максимально использует информацию, содержащуюся в выборке. Этот фильтр основан на использовании идей распознавания образов. В данной работе мы рассмотрим только принципы, заложенные в его основу.

Вначале все признаки приводятся к бинарному представлению. Выборка считается классифицированной на два класса. Первая выборка – анкеты одного пакета, вторая выборка – все остальные анкеты. Две выборки используются для построения эталонов классов (обучение). Затем производится распознавание наблюдений обоих выборок. По результатам распознавания рассчитывается количество ошибок распознавания. Далее процедура повторяется по схеме скользящего экзамена. Пакет анкет, обеспечивающий наименьшую ошибку распознавания, можно считать более обособленным и, следовательно, он рассматривается как аномалия, требующая содержательного анализа. При этом могут быть отброшены не только наблюдения (анкеты), принадлежащие этому пакету, но и прочие наблюдения, отнесенные к аномальной выборке в процессе распознавания.

Рассмотренный фильтр определим как **«Фильтр эталонных значений» (ФЭЗ)** и при рассмотрении пакетов – (**ФЭЗГ**) соответственно.

При содержательном анализе подозрительных данных, как правило, обращаются к первичному материалу (анкетам на бумажном носителе). При этом часто выбросы могут быть вызваны банальной ошибкой оператора или явной опиской респондента допущенной при заполнении анкеты, автоматически перенесенной оператором в компьютер. Такие ошибки легко устранимы. Явные выбросы улавливают практически все рассмотренные фильтры.

Все описанные выше фильтры обрабатывают ситуацию отсутствия значений в некоторых признаках. Для каждого фильтра разработан свой метод обработки ситуации отсутствия данных. Отсутствующие значения заменяются некоторыми их оценками. В связи с ограниченным объемом настоящей работы не представляется возможным рассмотреть эти методы более подробно.

Различные фильтры дают различные списки упорядоченных анкет. Однако явные выбросы улавливаются большинством фильтров. Для того чтобы сравнить результаты, полученные различными фильтрами, были взяты по двадцать более критичных анкет из каждого упорядоченного списка, составленного при обработке одних и тех же данных различными фильтрами. Для расчета опять были использованы данные анкеты по пляжно-оздоровительному отдыху. Степень совпадения списков можно представить в виде графа (*рис.3*). С целью обеспечения большей наглядности одиночные совпадения были исключены из графов. Большие совпадения на графиках отмечены жирной линией. На графе два метрических фильтра по Евклидову и Хэммингову расстоянию обозначены соответственно Ре ФМНЗ и Рх ФМНЗ. Степень совпадения результатов обработки пакетов анкет, представленных различными интервьюерами с помощью тех же фильтров, представлены на *рисунке 4*.

Модули, реализующие рассмотренные фильтры, снабжены специальными средствами проверки корректности данных. Диагностируется наличие пробелов, текстовых значений в числовых полях и т. п. Для отбора признаков, включаемых в фильтр, пользователю предъявляются для выбора только те признаки, которые могут быть обработаны данным фильтром. Программные модули, кроме значений критерия для элементов выборки, могут выводить информацию, которая может пояснить исследователю то, за счет чего было получено данное значение критерия. Эта информация может оказаться очень полезной при анализе выделенных программами многомерных выбросов.

Кроме серии многомерных статистических фильтров, разработанный комплекс программных средств, включает в свой состав еще ряд программ, предназначенных для обработки анкетных данных. Большинство программ имеют своей целью повышение качества анкетных данных. Примерами таких программ являются программы по восстановлению значений признаков, средства обработки открытых вопросов и другие. Программный комплекс реализован в форме надстройки к EXCEL – самому распространенному средству обработки данных.

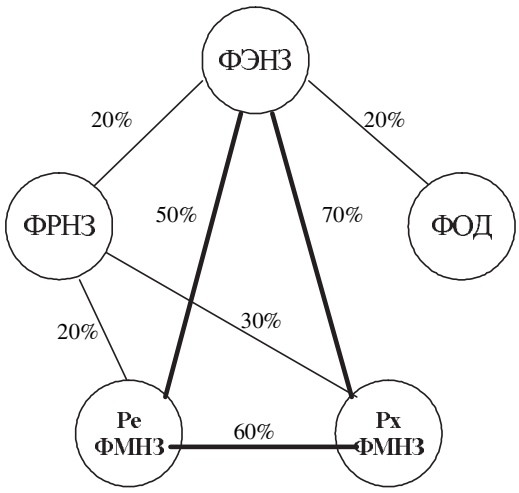


Рис. 3. Степень совпадения выявленных выбросов по отдельным анкетам

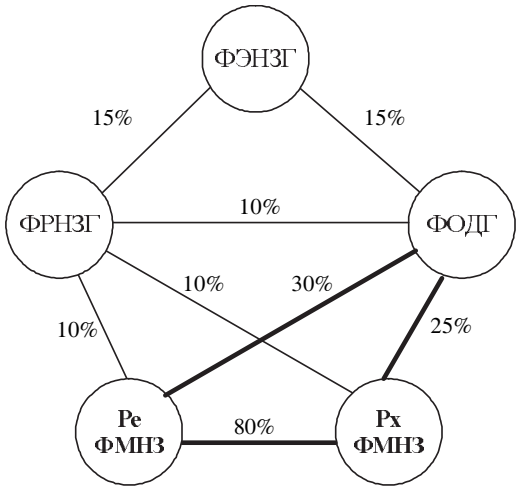


Рис. 4. Степень совпадения выявленных выбросов по пакетам анкет

После обработки анкетных данных с помощью программ повышения качества данных исследователь может далее воспользоваться любым другим программным средством, предназначенным для анализа статистических данных.

В настоящее время программы комплекса проходят опытную эксплуатацию. Они уже были использованы для анализа данных нескольких анкетных опросов и показали достаточно высокую эффективность.