Мартышенко Н.С., Мартышенко С.Н. Метод обнаружения ошибок в эмпирических данных // Известия вузов. Северо-Кавказский регион, 2008. – №1. С. 11–14.

2008

УДК 519.254

МЕТОД ОБНАРУЖЕНИЯ ОШИБОК В ЭМПИРИЧЕСКИХ ДАННЫХ

С.Н. Мартышенко, Н.С. Мартышенко

Владивостокский государственный университет экономики и сервиса

С развитием рыночной экономики в нашей стране все более возрастает конкуренция между предприятиями. Добиться успеха становится невозможным без широкого использования научных знаний по анализу управлению и рынков. Чем больше уровень конкуренции, тем большее внимание предприятия вынуждены уделять изучению мнений потребителей. 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. Поэтому все большее распространение имеют исследования, основанные на анкетных опросах населения. Современный подход к анализу анкетных данных требует использования статистических методов, позволяющих анализировать многомерные данные. Но методы многомерного анализа не находят широкого применения в практических исследованиях не потому, что исследователи не хотят их применять, а потому, что наблюдается большой недостаток в программных средствах на отечественном рынке программного обеспечения. Поэтому, большинство исследователей вынуждены ограничиваться первичной обработкой данных, используя доступные инструментальные средства. Вследствие этого огромное количество полезной информации, содержащейся в данных, остается не использованной.

Многомерные методы анализа позволяют более глубоко проникнуть в природу изучаемых объектов и явлений. Вместе с тем, многомерные методы предъявляют более высокие требования к качеству данных. Этап анализа качества данных приобретает особое значение и требует своего методического и программного обеспечения. Разнообразие типов данных усложняет задачу. Поэтому не может быть разработано одного универсального метода анализа качества данных, при-

годного на все случаи. Ошибки в данных анкетных опросов имеют очень сложную структуру. Для выявления различных видов ошибок требуется разные методы анализа.

Методы повышения качества данных неотрывно связаны с понятием грубой ошибки. Этому понятию невозможно дать однозначное формализованное определение. Поэтому попытаемся уточнить его через некоторые его свойства. Грубой ошибкой можно считать многомерное наблюдение, которое резко отличается на фоне всех остальных. Совокупность значений признаков можно считать грубой ошибкой, если они совместно воссоздают абсурдный, с содержательной точки зрения, объект или его поведение. При этом значения одномерных признаков могут быть вполне правдоподобными. Определить грань, за которой наступает абсурдность объекта, может только сам исследователь в процессе содержательного анализа многомерного объекта. Размытое определение грубой ошибки также приводит к разнообразию методов анализа качества данных.

На базе понятия грубой ошибки нами был разработан целый спектр алгоритмов, применимых для тех или иных групп признаков. Ряд методов обнаружения грубых ошибок в данных приведен в одной из работ авторов [1]. Предложенные алгоритмы работают по принципу многомерных фильтров, упорядочивающих данные в соответствии с некоторыми критериями. Окончательное решение по принятию решения об отнесении данных к ошибкам остается за исследователем. Однако для отдельных наблюдений, очень трудно принять решение относить ли их к ошибкам или считать правдоподобными. Такие наблюдения являются пограничными. Чтобы исключить неопределенность, в этих случаях желательно иметь некоторые дополнительные формализованные критерии. Эти критерии могут строиться на некоторых дополнительных предположениях о свойствах данных. Рассмотрим один из таких алгоритмов.

Данный алгоритм пригоден для анализа ошибок в многомерных данных, включающих только числовые признаки. При рассмотрении алгоритма значения признаков будем интерпретировать как точки многомерного пространства. Идею работы алгоритма рассмотрим на примере объектов, описываемых двумя призна-

ками (X_1, X_2). Первая часть алгоритма совпадает с известным алгоритмом классификации многомерных данных - КРАБ [3]. Поэтому эту часть алгоритма дадим в сокращенном виде.

Шаг 1. Рассчитывается матрица расстояний между всеми парами объектов.

Шаг 2. По многомерным точкам строится кратчайший незамкнутый путь или минимальное покрывающее дерево – каркас.

На этом, собственно, сходство с алгоритмом КРАБ и заканчивается.

Шаг 3. Производится классификация данных многомерной выборки. Для этого связи в графе последовательно разрываются, начиная с самой протяженной. Заранее количество классов, как в алгоритме КРАБ, не устанавливается. Количество классов определяется в результате дополнительного анализа классов, образующихся в результате деления каркаса. Образующиеся классы подразделяются на два типа. Классы первого типа – это “представительные классы”, то есть, объективно обособленные группы объектов. Классы второго типа – это “непредставительные классы” или “критические классы” - малочисленные классы, состоящие из особенных объектов, которые могут быть выбросами или образованы абсурдными данными (ошибками).

Для выделения типов классов используются два дополнительных параметра:

d_1 % - минимальный объем “представительного класса”, определенный в процентах от общего объема выборки;

d_2 % - максимальный суммарный объем “непредставительных классов” – порог.

Разбиение на классы заканчивается, когда суммарное количество объектов критических классов достигнет порогового значения d_2 %. Параметр d_2 % устанавливается исследователем и отражает его мнение о возможном проценте ошибок в данных. Этот процент может составлять от 5% до 10 % от общего объема выборки. Полученный результат можно представить графически (рис.1.).

На рис. 1 представлено два “представительных” класса (номера 1, 2) и четыре “непредставительных” класса (номера 3-6).

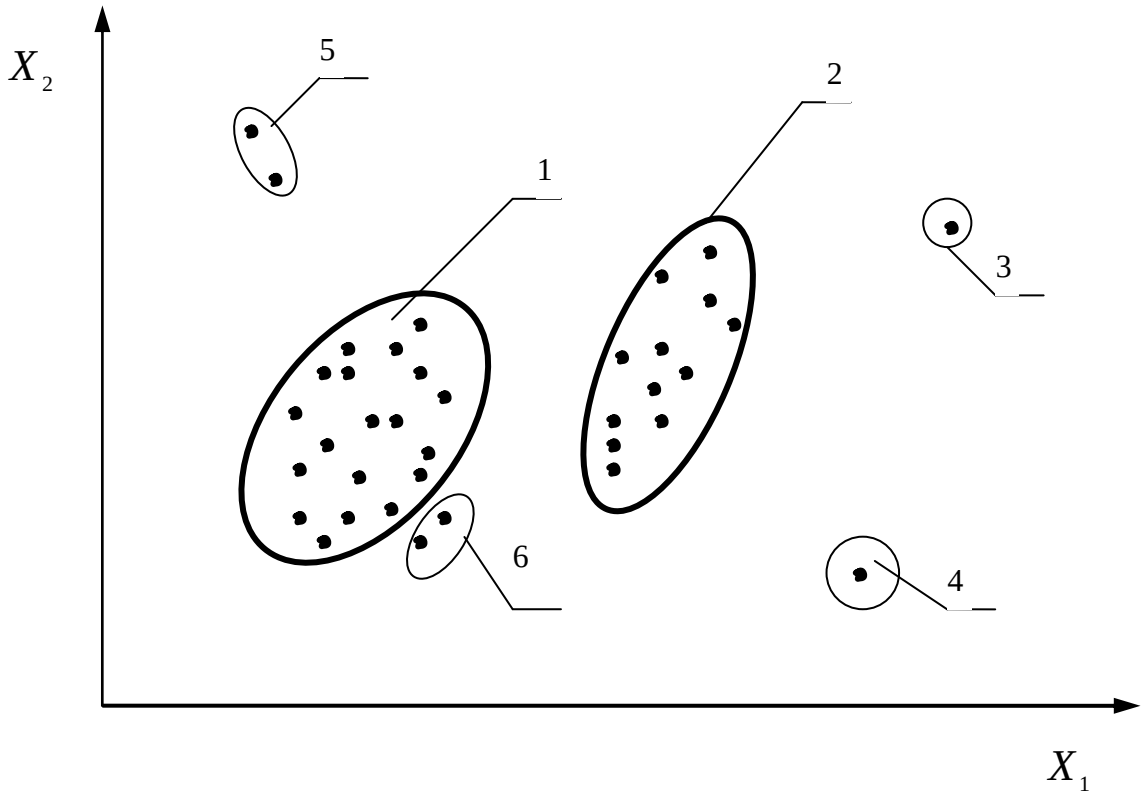


Рисунок 1. Два типа классов: “представительные” и “непредставительные”

Шаг 4. На этом шаге анализируются объекты, составляющие “непредставительные” классы. В результате анализа определяется, являются ли объекты “непредставительных” классов действительно выбросами или объекты были отделены от одного из “представительных” классов случайно и должны быть присоединены к такому классу.

Основная идея выделения ошибки состоит в формальном описании некоторых свойств представительных классов. Если при присоединении к такому классу нового объекта свойства класса сохраняются, то объект можно объединить с существующим классом, иначе такой объект можно считать выбросом.

В рассматриваемом алгоритме предлагается описывать свойство класса на основе замкнутой выпуклой оболочки, построенной по точкам образующим класс. Естественно в качестве основного класса рассматривать ближайший представительный класс. Известно несколько алгоритмов построения выпуклой оболочки по группе точек [2]. Нами была использована сходная процедура. Рассмотрим выпуклый контур представительного класса 1 и критический класс 6 (рис. 2).

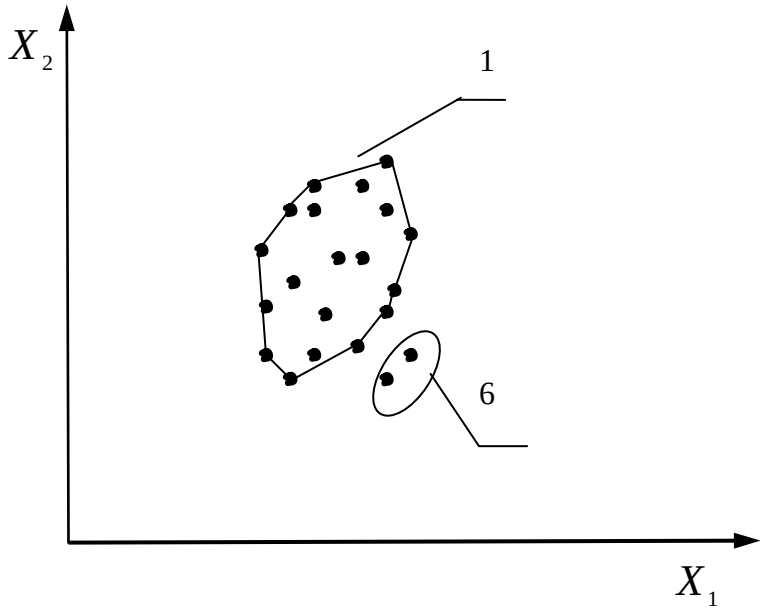


Рисунок 2. Выпуклая оболочка представительного класса номер 1

Для характеристики оболочки класса введем новые понятия. Рассмотрим три смежные вершины контура (рис. 3).

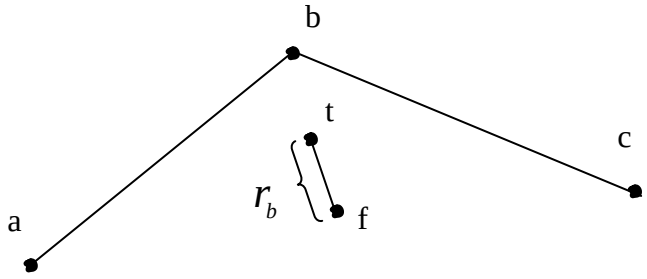


Рисунок 3. Фрагмент выпуклого контура класса

Рассчитаем для промежуточной точки b некоторое значение r_b, которое назовем разнообразием точки b выпуклого контура класса. Показатель r_b - это расстояние между двумя точками $t(x_1^t, x_2^t)$ и $f(x_1^f, x_2^f)$, координаты которых рассчитаем по формулам (1-2):

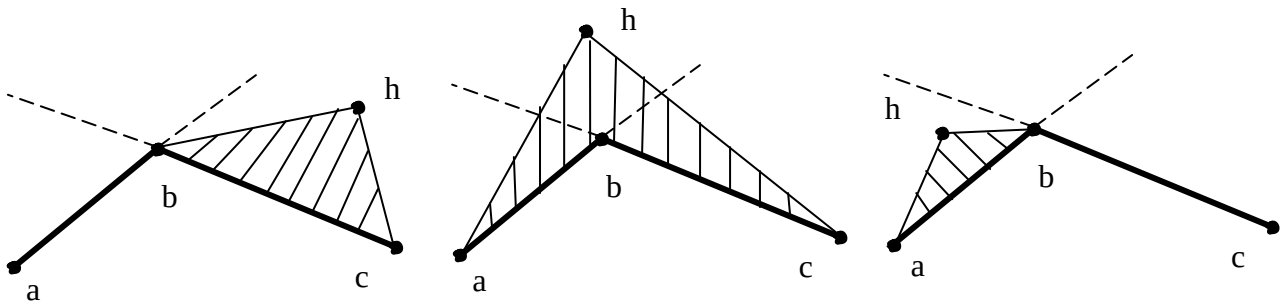
$$x_1^t = \frac{x_1^a + x_1^b + x_1^c}{3}, \quad x_2^t = \frac{x_2^a + x_2^b + x_2^c}{3}, \tag{1}$$

$$x_1^f = \frac{x_1^a + x_1^c}{3}, \quad x_2^f = \frac{x_2^a + x_2^c}{3}. \tag{2}$$

То есть, точки $t(x_1^t, x_2^t)$ и $f(x_1^f, x_2^f)$ ассоциируются с центром тяжести, оцененным по трем точкам (*a*,*b*,*c*) и двум точкам (*a*,*c*) соответственно. Для характеристики всего контура будем использовать показатель ρ -разнообразия контура:

$$\rho = \max_{\substack{\text{по всем } t \text{ точк} \\ \text{контура}}} (r) \tag{3}$$

Для точки *h* из критического класса, расположенной ближе всего к одному из представительных классов, рассчитывается характеристика разнообразия исходя из предположения принадлежности точки *h* представительному классу. Предположим, что ближайшей точкой контура класса к точке *h* является точка *b*. В зависимости от расположения точки *h* относительно точек контура *a*,*b* и *c* возможны три варианта расчета разнообразия точки *h* (рис. 4). На этом рисунке контур с участием точки *h* изображен более тонкой линией, чем основной контур.



1) 2) 3)

Рисунок 4. Три варианта расположения внешней точки *h*

Если значение разнообразия r_h точки *h* окажется меньше либо равно разнообразию всего контура ρ, то к представительному классу присоединяются все точки критического класса, включающего точку *h*, иначе все точки критического класса идентифицируются как выбросы или ошибки.

Рассмотренный алгоритм представляет собой только схему расчета. Для того чтобы он стал пригоден для работы с реальными данными, в него был внесен ряд дополнений.

Для практического использования алгоритма очень важно время выполнения расчетов. Реальные данные могут содержать тысячи наблюдений и десятки признаков. И если бы данные не содержали совпадающих значений, то расчет мог

бы занять слишком много времени. В реальных данных количество различных комбинаций значений признаков ограничено. Так при двух признаках количество различных комбинаций лежит в диапазоне от 50 до 100. С ростом количества признаков количество комбинаций быстро растет. В случае повторяющихся комбинаций простое сжатие списка данных с учетом кратности существенно сокращает время на выполнение расчета.

При количестве признаков выше двух теоретически возможно было бы строить многомерные оболочки классов. Однако это, с одной стороны, очень усложнило бы расчет, с другой стороны, существенно увеличило бы время выполнения расчета. Поэтому при анализе пространства признаков больше двух нами рассматриваются проекции объектов на плоскости возможных сочетаний пар признаков.

При построении контуров классов была учтена еще одна особенность, присущая реальным данным. Практика показывает, что при ответах на вопросы, выражающиеся числом, респонденты чаще всего используют круглые или кратные числа. Такую ситуацию для анкетных опросов можно признать вполне нормальной, поскольку от респондентов требуется дать численную оценку некоторых средних величин. В этом случае, при проекции классов на плоскость двух признаков граница может включать повторяющиеся значения по одному из признаков (рис. 5). Эффективность работы алгоритма повышается, если в контур включаются все промежуточные точки отрезка (a,b).

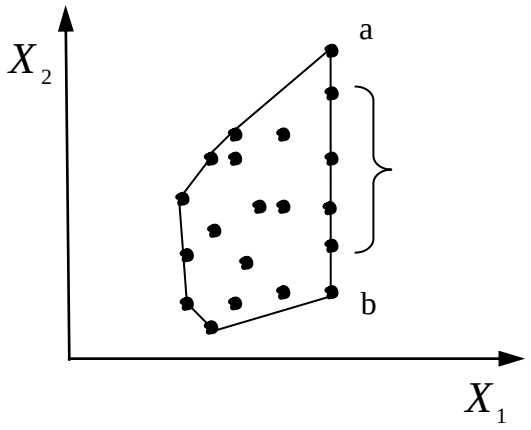


Рисунок 5. Промежуточные точки контура

Алгоритм был апробирован на модельных данных и на данных нескольких массовых анкетных опросов. Для моделирования данных использовалась модель многомерной нормальной выборки [4]. Результаты даже превзошли ожидания. Это можно объяснить тем, что нормальные данные очень хорошо описываются выпуклыми оболочками.

При работе с реальными данными иногда возникают неопределенные ситуации, связанные с наличием разряженного фона. Но это больше отражается на задаче классификации. Здесь мы ставим несколько иную цель - поиск выбросов или ошибок. Поэтому иногда можно мериться с неопределенностью при формировании больших классов. Мы уже говорили о невозможности построения абсолютного алгоритма выявления ошибок. В одних случаях будет работать лучше один алгоритм в других - другой. Поэтому для анализа ошибок необходимо использовать ряд алгоритмов. При работе с различными программами обнаружения ошибок очень важен опыт исследователя.

Работа по исследованию возможностей и совершенствованию алгоритма продолжается. В дальнейшем для характеристики свойств классов предполагается использовать и другие свойства описания единообразия классов. Для областей, которые плохо описываются выпуклыми оболочками, можно разбивать область на несколько выпуклых областей. При очень большом количестве объектов для ускорения работы алгоритма можно оценивать разнообразие по части контура, прилегающей к критической точке.

Кроме алгоритмов выделения ошибок в числовых данных, мы разрабатываем и алгоритмы обнаружения ошибок в данных, представленных качественными признаками. В настоящее время нами разработан ряд простых, но эффективных процедур [5].

Рассмотренный алгоритм приобретает наибольшую практическую ценность при совместном использовании с другими средствами анализа анкетных данных. Такие средства мы объединили в единый программный комплекс [6]. Этот комплекс постоянно дополняется новыми модулями. Программы комплекса исполь-

зуются для обработки анкетных опросов, проводимых на кафедре маркетинга и коммерции ВГУЭС.

Литература

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

2. Загоруйко, Николай Григорьевич Прикладные методы анализа данных и знаний / Н. Г. Загоруйко – Новосибирск Изд-во Ин-та математики СО РАН, 1999 - 270 с.

3. Препарата Ф., Шеймос М. Вычислительная геометрия: Введение / Пер. с англ. – М.: Мир, 1989. – 478 с.

4. Мартышенко Н.С., Мартышенко С.Н., Кустов Д.А. Моделирование многомерных данных // Техника и технологии. — 2007. — №4(5). С. 468–478.

5. Мартышенко Н.С., Власенко А.А. Устранение ошибок в данных маркетинговых исследований с помощью логических фильтров // Наука и образование – 2007. Материалы V Международной научно-практической конференции. Том 7 Экономические науки. — Днепропетровск. 2007. — С. 76-78.

6. Мартышенко Н.С., Мартышенко С.Н., Кустов Д.А. Совершенствование математического и программного обеспечения обработки первичных данных в экономических и социологических исследований // Вестник Тихоокеанского государственного экономического университета. — 2006. — №2. С. 91–103.