Известия высших учебных заведений. Поволжский регион. – 2006. – №6(27). С. 101–110.

ПРИМЕНЕНИЕ АЛГОРИТМОВ МНОГОМЕРНОЙ КЛАССИФИКАЦИИ И РАСПОЗНАВАНИЯ ОБРАЗОВ В РЕШЕНИИ ЗАДАЧ АНАЛИЗА АНКЕТНЫХ ДАННЫХ

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

С развитием экономики нашей страны все более возрастает конкуренция между предприятиями. Добиться успеха становится невозможным без широкого использования научных знаний управления и анализа рынка. Все большее внимание предприятия вынуждены уделять изучению мнений потребителей.

Основным источником первичных данных отражающих мнения потребителей являются анкетные опросы потребителей. В связи с этим наблюдается рост потребности в новых методах и методиках обработки таких данных.

Анкетные данные представляют собой выборки многомерных данных различной природы. Одним из наиболее эффективных средств анализа многомерных данных, является теория многомерной классификации и распознавания образов. Спектр задач анализа анкетных данных, которые могут быть решены с использованием методов многомерной классификации и распознавания образов достаточно широк. С помощью таких методов можно решать следующие задачи:

1) задача выявления выбросов или грубых ошибок;

2) задача восстановления данных;

3) задача выделения однородных групп объектов (классификация);

4) задача прогнозирования признаков (распознавание по обучающей выборке).

На сегодняшний день методы многомерной классификации используются только для решения третьей задачи и то в очень ограниченных масштабах. Возможности решения третьей задачи с помощью стандартных пакетов по статистике, включающих модули классификации и распознавания, на самом деле весьма ограничены. Ограничение возникает из-за того, что распространенные пакеты позволяют решать задачу классификации с признаками, измеренными в относительной шкале, а анкеты порождают большое количество нечисловых признаков. Для решения задач классификации анкетных данных необходимы алгоритмы, которые могли бы рабо-

тать с признаками, измеренными в различных шкалах. Таким свойством обладает некоторые непараметрические методы распознавания образов.

В настоящей работе предлагается использовать для решения четырех задач, приведенных выше, идеи метода интегральной диагностики многомерных систем. Основные теоретические положения, заложенные в основу метода интегральной диагностики многомерных систем, изложены в работах Б.И. Адасовского [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. Он был использован в основном для разработки технических систем.

Для каждой из 4-х задач анализа анкетных данных нами разрабатывался собственный алгоритм и компьютерная программа. В плане теории они имеют некоторое общее ядро. Поэтому вначале мы рассмотрим это ядро в форме обобщенного алгоритма, а потом представим специфику его использования при решении четырех задач анализа данных анкетных опросов.

Алгоритм представим как последовательность вычисления элементов основных информационных массивов данных (рис. 1).

Исходной информацией для расчета служит многомерная выборка, которую можно представить как матрицу A . Количество строк матрицы A равно количеству наблюдений выборки n_0, первые m столбцов содержат значения признаков по всем наблюдениям, последний столбец задает номер класса, к которому отнесено наблюдение, в соответствии с установленным правилом классификации. Другими словами, матрица A представляет собой ничто иное, как обучающую выборку. Исследователем должно быть определено единое для всех признаков количество интервалов дискретизации L. Исходная многомерная обучающая выборка, в соответствии с введенным количеством дискретов, преобразуется в матрицу дискретных значений $Б$, содержащую n_0 строк и $m \times L$ столбцов.

Считается, что в исходном состоянии все элементы матрицы B_1 имеют нулевые значения. Для каждого элемента первого столбца матрицы $Б$ проделывается следующая операция. Определяется соответствующий номер класса, который задает

строку матрицы B_1, а значение этого элемента (номер дискрета) задает столбец матрицы B_1 (строка и столбец определяют элемент матрицы B_1), к которому прибавляется единица.

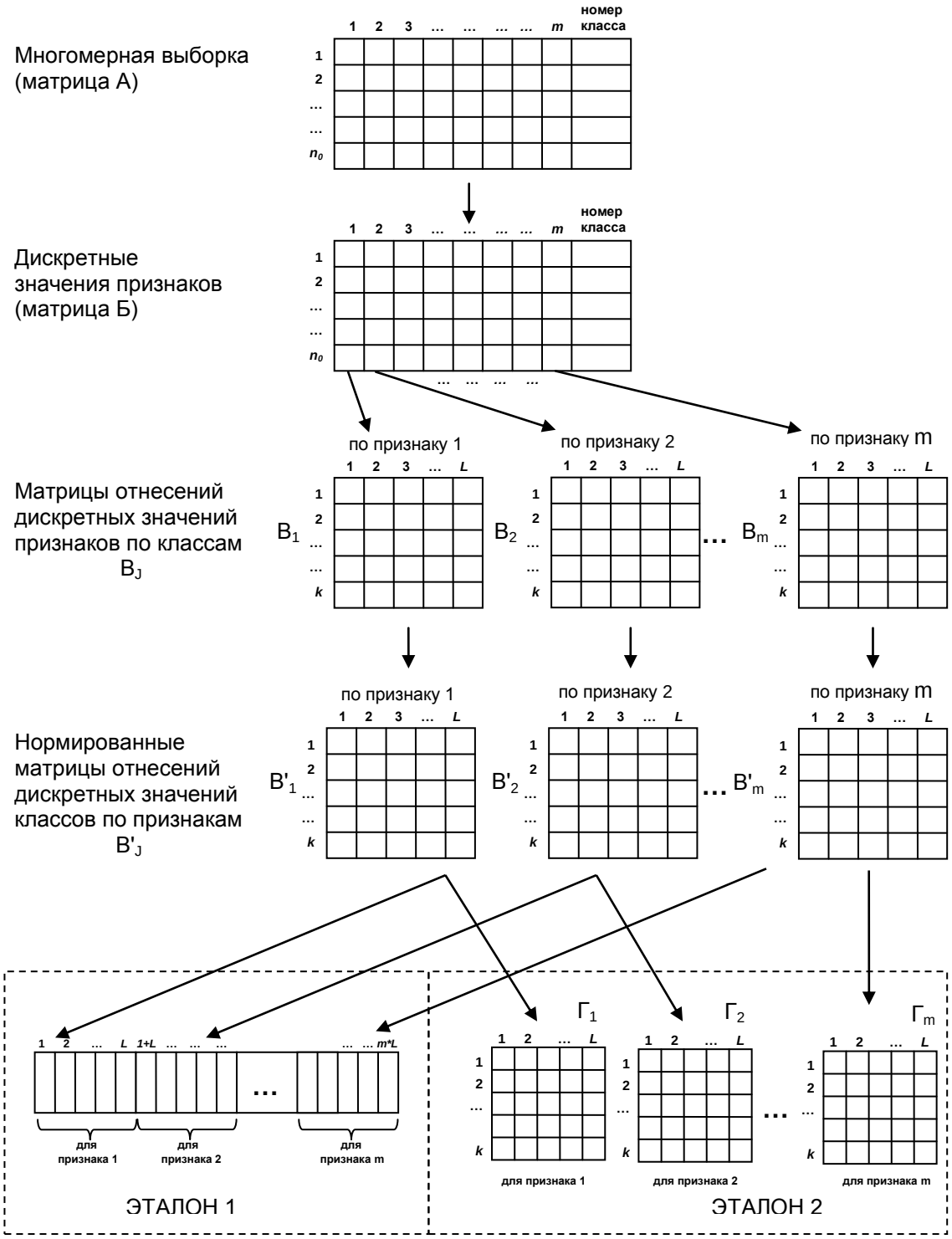

Рис. 1. Структура основных информационных массивов, используемых программами анализа анкетных данных

Таким образом, сумма элементов каждой из матриц $B_1, B_2, \ldots, B_m$ будет равна количеству наблюдений многомерной выборки n_0 , Матрицы $B'_1, B'_2, \ldots, B'_m$ получа-

ются путем нормирования по строкам соответствующих матриц $B_1,B_2,\ldots,B_m$. Нормирование производится делением строк матриц $B_1,B_2,\ldots,B_m$. на количество наблюдений в соответствующих классах $n_1,n_2,\ldots,n_k$. Информация, хранящаяся в этих матрицах используется для расчета конечных результатов: эталонов классов. Для расчета эталонов классов могут использоваться две схемы.

Рассмотрим первую схему. Эталоны классов хранятся в виде одномерного массива, содержащего $m \times L$ элементов. Первые L элементов рассчитываются на основании матрицы B'_1, следующие L элементов - на основании матрицы B'_2, и т.д. Каждому столбцу всех матриц $B_1,B_2,\ldots,B_m$ соответствует элемент массива, задающего эталоны классов. Эти элементы рассчитываются следующим образом: в результате просмотра соответствующего столбца определяется номер строки, содержащей максимальный элемент (номер класса), который присваивается соответствующему элементу эталона.

Рассмотрим схему распознавания произвольного наблюдения рабочей выборки (для которого номер класса не известен) с использованием полученного эталона (рис. 2).

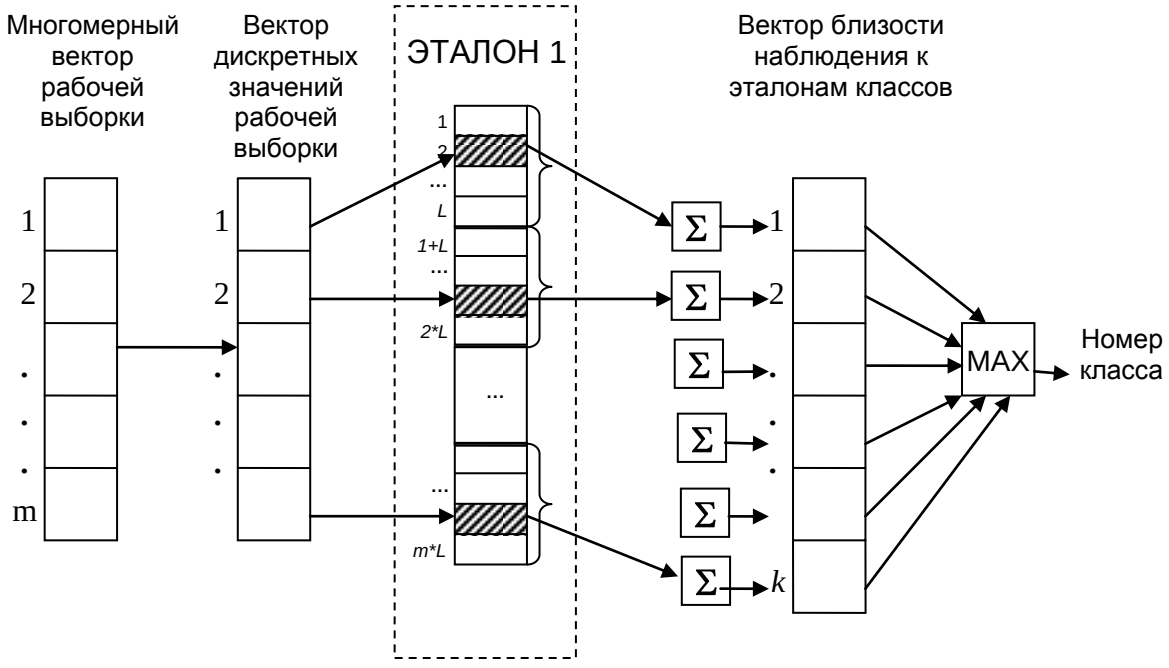

Рис. 2 Первая схема распознавания произвольного наблюдения рабочей выборки по эталонам классов

В начале производится дискретизация значений признаков по дискретам, выбранным для формирования эталона. В результате получаем вектор дискретных значений наблюдения рабочей выборки. По номерам дискретов этого вектора определяем номера классов из эталонов. Количество попаданий в классы записывается в вектор оценок близости наблюдения рабочей выборки к эталону. Этот вектор выполняет роль дискриминантной функции. Наблюдению рабочей выборки присваивается номер класса, имеющего наивысшую оценку близости к эталону класса.

Рассмотрим вторую схему расчета эталонов классов. Эталоны классов хранятся в виде m матриц $\Gamma_1,\Gamma_2,\ldots,\Gamma_m$, каждая из которых содержит k строк и l столбцов. Матрицы $\Gamma_1,\Gamma_2,\ldots,\Gamma_m$ рассчитываются на основании матриц $B_1,B_2,\ldots,B_m$. Столбцы $\Gamma_1,\Gamma_2,\ldots,\Gamma_m$ получаются в результате нормирования соответствующих столбцов матриц $B_1,B_2,\ldots,B_m$. Нормирование производится в соответствии с формулой:

$$\Gamma_{ij\psi} = \frac{B_{ji\psi}}{k}, \qquad (1)$$

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

i - номер класса, $i = 1, 2, ..., k$;

ψ - номер дискрета, $\psi = 1, 2, ..., L$.

В результате нормирования сумма элементов каждого столбца матриц $\Gamma_1,\Gamma_2,\ldots,\Gamma_m$ становятся равны 1. Рассмотрим схему распознавания произвольного наблюдения рабочей выборки с использованием полученного эталона (рис. 3).

Дискретизация признаков наблюдения рабочей выборки производится аналогично первой схеме. По номерам дискретов вектора дискретных значений признаков наблюдения рабочей выборки определяем номера столбцов в матрицах $\Gamma_1,\Gamma_2,\ldots,\Gamma_m$, задающих эталон. В результате поэлементного суммирования выбранных столбцов получаем вектор оценок близости наблюдений рабочей выборки к эталону. Разделив вектор оценок близости к эталону на число признаков m , получим вектор оценок вероятности принадлежности наблюдения рабочей выборки к классам $\Omega_1,\Omega_2,\ldots,\Omega_k$.

К числу недостатков изложенного выше алгоритма можно отнести то, что разные по информативности признаки играют в нем одинаковую роль при распознава-

нии. Однако этот вопрос требует дальнейших теоретических и экспериментальных исследований.

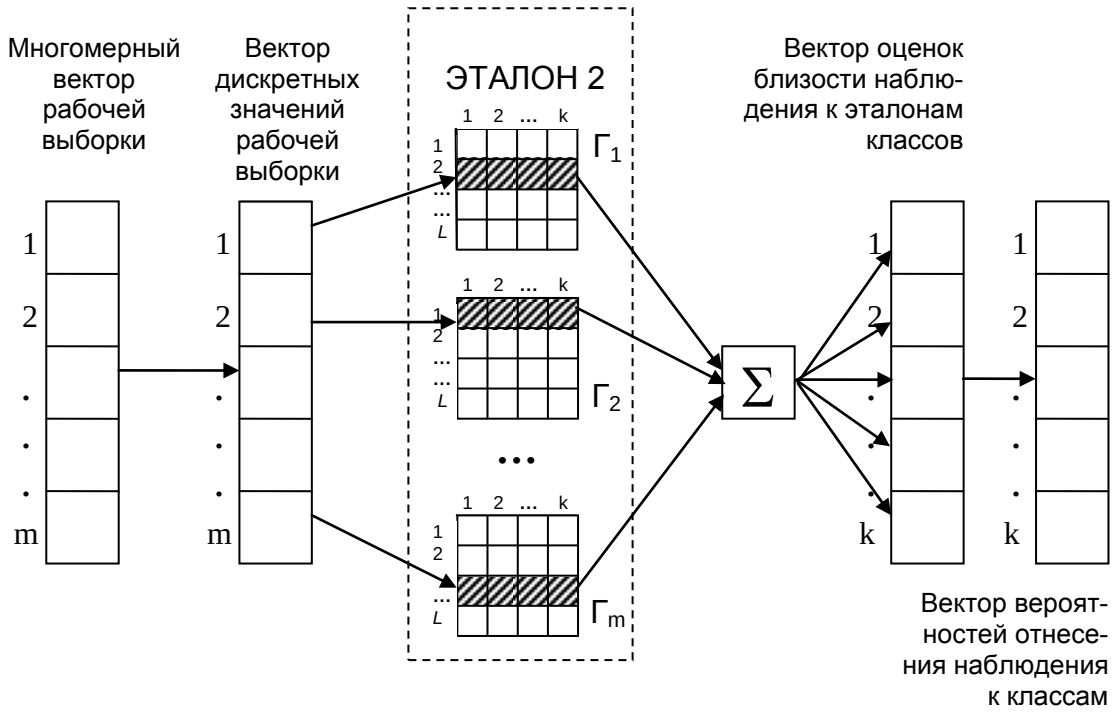

Рис. 3. Вторая схема распознавания наблюдения рабочей выборки по эталонам классов

Рассмотрим приложение данного алгоритма для решения четырех приведенных выше задач анализа анкетных данных. Чтобы использования алгоритма для решения каждой задачи были разработаны специальные программные модули, которых включены в состав разработанного нами пакета анализа анкетных данных предназначенного для работы в среде EXCEL [3,4].

Первая задача – выявление выбросов. Программный модуль, предназначенный для решения этой задачи, работает с признаками, представленными в номинальной или порядковой шкале. Для того чтобы включить в состав признаков и признаки, представленные в относительной шкале, эти признаки необходимо предварительно преобразовать к порядковой шкале. Преобразование производится с помощью специальной программы дискретизации признаков.

В отличие от рассмотренного обобщенного алгоритма, признаки могут иметь различное количество уникальных значений (дискрет). Обозначим количество дискрет j - го признака через L_j , j=1,2,….m, где m –количество признаков. Признаки мо-

гут принимать не только кодированные числовые значения, но и текстовые. Некоторые значения признаков могут отсутствовать.

Работу фильтра рассмотрим на примере выявления резко отличающихся пакетов анкет, собранных различными интервьюерами.

Обучающая выборка для этой задачи имеет всего два класса. В первый класс входят данные одного интервьюера. Во второй класс все остальные данные. По данным обучающей выборки формируется эталон классов. Затем все наблюдения первого класса подвергаются распознаванию по первой схеме (рис. 2). Отличие процедуры распознавания заключается в том, что признакам присваиваются различные веса q_j:

$$q_j = \frac{L_j}{\sum_{i=1}^{m} L_j} \tag{2}$$

Если один из признаков для распознаваемого объекта не определен, то появляется неопределенность отнесения наблюдения по этому признаку к одному из двух классов. В этом случае оба класса получают одинаковое приращение дискриминантной функции.

Предположим, что при распознавании n_j объектов j-го интервьюера n_j^1 - объектов отнесено к классу 1, то есть к своему классу, а n_j^2 - объектов отнесены к другому классу, который составлен из всех элементов выборки за исключением объектов j-го интервьюера. Тогда каждому интервьюеру можно поставить в соответствие некоторый критерий:

$$\omega_j = n_j^1 / n_j, \tag{3}$$

Который имеет смысл доли правильно распознанных объектов j-го интервьюера. Этот показатель принимает значение от нуля до единицы. Чем ближе этот показатель к единице, тем более анкеты интервьюера отличаются от всех остальных. Поэтому этот критерий вполне может быть использован для выявления резко отличающихся пакетов (выбросов).

Если использовать алгоритм распознавания для анализа выбросов не по пакетам, а по отдельным наблюдениям, то первый класс будет состоять только из одного наблюдения и поэтому надо применять вторую схему распознавания (рис.3). В этом случае критерием отличия объектов можно использовать значение дискриминантной функции первого класса. Чем выше это значение, тем объект более уникален. Применение этого подхода для выявления выбросов имеет смысл только при достаточно большом количестве признаков (не менее 5).

Вторая задача - задача восстановления данных. 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.

Третья задача - задача выделения однородных групп объектов (классификация). Эта задача имеет одно из центральных мест в маркетинговых и социологических исследований. В маркетинге эта задача известна, как задача сегментирования.

Рассмотрим идею использования алгоритма для решения задачи классификации. Идея классификации близка известному алгоритму *k*-средних. Но она имеет и ряд принципиальных отличий.

Алгоритм классификации строится в два этапа. Работу алгоритма продемонстрируем на гипотетическом примере (рис. 4). На первом этапе случайным образом выбираются *k* объектов из n_0 объектов выборки. Эти элементы составляют обучающую выборку на первом этапе. Тогда контрольная выборка будет состоять из всех оставшихся элементов выборки.

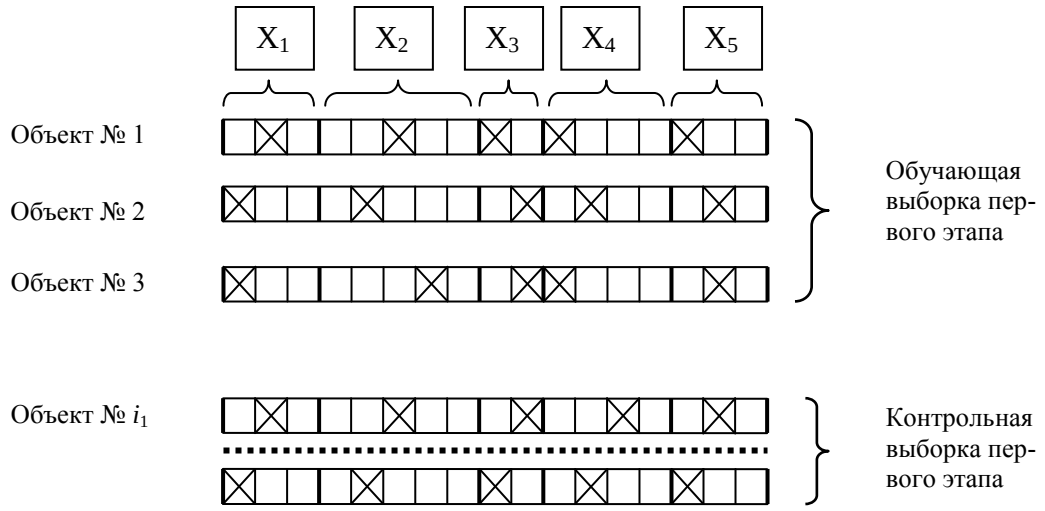

Рис. 4. Гипотетический пример выборки из пяти признаков

В примере k=3. Количество признаков участвующих в классификации m=5. Признаки на схеме (рис. 4) обозначены, как X_1,X_2,X_3,X_4,X_5. Количество дискрет по каждому признаку можно описать множеством {3;5;2;4;3}. Тогда веса признаков q_j будут описываться множеством {0,18; 0,29; 0,12; 0,24; 0,18}. Произведем распознавание всех остальных элементов выборки, сравнивая их с k объектами обучающей выборки.

В примере производится распознавание двух произвольных элементов контрольной выборки с номерами i_1 и i_2 соответственно. Распознавание производится путем подсчета совпадающих дискрет для объектов обучающей выборки и распознаваемого объекта. Затем вычисляется взвешенная сумма совпадений. Эта сумма служит для определения принадлежности распознаваемого объекта к классу обучающей выборки.

Если в обучающей выборке не встретился дискрет, указанный в объекте распознавания, то всем классам присваивается равный вес $1/k$. Если дискрет встречается сразу в нескольких объектах обучающей выборки, то веса классов распределяются равномерно по всем таким классам. На первом этапе нецелесообразно включать в контрольную выборку объекты с пропущенными значениями признаков.

Взвешенные суммы служат для определения номера класса для объектов обучающей выборки. Так для объекта с номером i_1 распределение взвешенных сумм будет следующее {0,412; 0,235; 0,235}, а для объекта с номером i_2 {0,353; 0,441;

0,088}. Поэтому объект с номером i_1 мы относим к первому классу, а объект с номером i_2 относим ко второму классу. Если значение двух наибольших значений дискриминантной функции отличается менее чем на 20 %, то такие объекты лучше оставить нераспознанными.

Первый этап заканчивается после распознавания всех элементов контрольной выборки. При этом часть элементов могут остаться нераспознанными. Распознанные элементы составят новую контрольную выборку.

Второй этап выполняется как ряд повторяющихся итераций. Созданная на первом этапе обучающая выборка используется для построения эталонов классов. Затем все данные подвергаются распознаванию по первой схеме. По распознанным объектам строится новая обучающая выборка. Пересчет эталонов повторяется до тех пор, пока классы не стабилизируются, то есть для всех элементов выборки номера классов на очередном шаге не изменятся.

Поскольку на первом этапе для формирования эталонов классов используется случайный отбор объектов, то всю процедуру целесообразно повторить 5-10 раз. Если при классификации не удается получить соизмеримые по объему классы, то необходимо определять новую систему признаков классификации.

Четвертая задача - задача прогнозирования классов (распознавание по обучающей выборке). Эта задача отличается предыдущей, что предполагает наличие двух различных по качеству выборок. Предположим, мы провели опрос некоторой группы респондентов, который сопровождался дополнительным обсуждением с ними проблемы опроса. В результате бесед с респондентами исследователь самостоятельно разделил респондентов по классам или группам.

Однако, исследователь не в состоянии охватить большое количество респондентов глубокому анализу и соблюсти репрезентативность выборки. Зато он может по небольшой группе установить признаки, которые хорошо разделяют классы. Затем исследователь может использовать эти признаки для организации массового опроса. А для определения (прогнозирования) классов второй выборки он может использовать первую выборку.

Все четыре рассмотренные задачи основаны на применении общего подхода, но схема расчета и структура используемых данных у них различны. Отличаются и результаты решения задач. Поэтому для решения этих задач было разработано четыре различных программных модуля. Эффективность работы программных модулей была исследована на модельных данных. Программы были апробированы при обработке ряда реальных анкет.

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.