УДК 519.25

МНОГОМОДЕЛЬНЫЙ ПОДХОД К МАССОВОЙ ОЦЕНКЕ МНОГОПАРАМЕТРИЧЕСКИХ ОБЪЕКТОВ

Е.К. Корноушенко

На примере массовой (регрессионной) оценки объектов недвижимости описывается новый подход к массовой оценке многопараметрических объектов. В отличие от традиционного подхода с использованием одной модели для оценки массива рассматриваемых объектов в новом подходе предлагается использовать несколько моделей. Оцениваемые объекты при этом классифицируются, и объекты каждого класса оцениваются соответствующей моделью, что существенно повышает правдоподобность получаемых оценок. Изложение иллюстрируется практическим примером.

Ключевые слова: массовая оценка, регрессионная модель, непрерывная, категориальная переменная, алгоритм классификации, дисперсионный анализ, линейный дискриминантный анализ Фишера.

ВВЕДЕНИЕ

Настоящая работа является, по существу, продолжением работы [1], в которой сформулирована задача массовой оценки с учетом наличия в рыночной информации о стоимости объектов ненаблюдаемой составляющей. Наличие такой составляющей приводит к тому, что при построении по рыночной информации модели, используемой для массовой оценки, приходится жестко фильтровать (отбраковывать) объекты с рыночной ценой для обеспечения требуемого качества оценки строящейся модели. С целью уменьшения числа отбракованных объектов и более полного использования рыночной информации о стоимости объектов в работе [1] описана специальная итерационная процедура построения моделей оценки и предложено для улучшения качества массовой оценки использовать не одну, а две модели оценки. Оцениваемые объекты вначале классифицировались на два класса с помощью предложенного алгоритма классификации [1], а затем оценивались с помощью модели того класса, к которому принадлежит оцениваемый объект.

В настоящей работе процедура построения моделей оценки, изложенной в работе [1], расширена на случай более двух моделей. Увеличение числа моделей оценки способствует лучшему учету исходной рыночной информации об объектах на оцениваемой некоторой территории и позволяет получить более правдоподобную картину распределения стоимости объектов на этой территории. Переход к большему числу моделей потребовал модификации алгоритма классификации на случай, когда классов больше двух. Проведенная модификация интересна тем, что на выходе алгоритма информация о каждом из классов взвешивается с помощью линейной регрессионной модели, а в роли классифицирующего правила выступает процедура округления значений непрерывного выхода регрессионной модели до целочисленных значений (номеров классов). Изложение иллюстрируется практическим примером массовой оценки удельной стоимости квартир в многоэтажных домах в разных районах г. Сочи. Проведенное в данном примере сравнение классифицирующей способности модифицированного алгоритма и соответствующей логистической модели показало существенное превосходство алгоритма. Особо подчеркивается наличие этапа верификации оценки объектов большого массива, не содержащих информации об их рыночной стоимости. Классификация таких объектов и верификация их оценки отличают предлагаемый подход от известных подходов к массовой оценке.

1. ИТЕРАЦОННАЯ ПРОЦЕДУРА ПОСТРОЕНИЯ МОДЕЛЕЙ ДЛЯ МАССОВОЙ ОЦЕНКИ

Исходной информацией для построения моделей оценки служит *рыночная выборка* (РВ), полученная после обработки первоначальной выборки

с рыночной информацией (удаления выбросов в значениях факторов и/или стоимости объектов, доопределения недостающих значений и др.). Путем анализа парных коэффициентов корреляции каждого из факторов объектов из РВ с их стоимостью выделяется совокупность, скажем, из m факторов стоимости, одна и та же для всех строящихся далее моделей оценки. Как и в работе [1], критерием качества оценки отдельного объекта принята *относительная погрешность оценки* (ОПО), обозначаемая как δ и определяемая для i-го объекта по формуле $\delta_i = \dfrac{\left|Y_i - \hat{Y}_i\right|}{Y_i}$, где Y_i — рыночная стоимость, а $\hat{Y}_i$ — модельная стоимость i-го объекта. При этом задается максимальное допустимое значение δ_{max} ОПО для оцениваемых объектов. По исходной РВ строится начальная модель $M_{0нач}$, которая разбивает РВ на два класса: класс C_0^0 «пригодных» объектов, ОПО которых моделью $M_{0нач}$ не превышает заданного значения δ_{max}, и класс C_0^1 «забракованных» объектов. На множестве C_0^0 строится модель M_1 с помощью которой снова оцениваются все объекты РВ. При этом ОПО объектов из C_0^0, оцениваемых моделью M_1, может быть как меньше δ_{max}, так и больше δ_{max}. Объекты с ОПО, большей δ_{max}, переносятся из C_0^0 в C_0^1, а объекты из C_0^1 с ОПО, меньшей δ_{max}, переносятся из C_0^1 в C_0^0. В итоге модель M_1 разбивает РВ на два класса: класс C_1^0 «пригодных» для модели M_1 объектов с ОПО $\le \delta_{max}$, и класс C_1^1 «забракованных» моделью M_1 объектов с ОПО, большей δ_{max}. Так заканчивается первый шаг итерационного процесса построения моделей, представленного в табл. 1.

В каждой итерации верхний нулевой индекс у множеств обозначает «пригодные» объекты (относительно модели, построенной на данной итерации), а единичный индекс — «забракованные». Данный итерационный процесс можно представить как функционирование конечного автомата, состояние которого на каждом такте итерации определяется парой множеств «пригодных» и «забракованных» объектов, а функция переходов зависит от построенной текущей модели. В силу конечности РВ множество состояний такого автомата конечно, т. е. в процессе функционирования автомат должен «зациклиться», вернуться в некоторое ранее пройденное состояние. Но тогда и модель, построенная в этом состоянии, будет, очевидно, совпадать с моделью, ранее построенной в этом ранее пройденном состоянии, т. е. траектория автомата начнет повторяться (заметим, что в приводимом далее примере автомат впервые приходит в устойчивое состояние на 10-й итерации). В итоге все объекты исходной РВ разбиваются на два класса: класс K_0 объектов, на которых построена результирующая модель $M_{0рез}$, удовлетворяющая заданным требованиям по качеству оценки, и класс K_1 «забракованных» объектов.

В принципе, для достаточно мощного[1] класса K_1 можно построить модель $M_{1нач}$, разбивающую его

Таблица 1

Псевдокод итерационного процесса построения по РВ моделей стребуемым качеством оценки

Задание начальных условий: РВ, m, δ_{max}
Построение модели $M_{0нач}$ на РВ: классы C_0^0 и C_0^1 i-я итерация: классы C_{i-1}^0, C_{i-1}^1
Построение модели M_i на C_{i-1}^0
Удаление из C_{i-1}^0 объектов O_i^0, для которых $\delta > \delta_{max}$
Добавление в C_{i-1}^0 объектов $O_i^1 \in C_{i-1}^1$, для которых $\delta < \delta_{max}$
Формирование классов $C_i^0 = O_i^1 \cup C_{i-1}^0 \setminus O_i^0$, $C_i^1 = O_i^0 \cup C_{i-1}^1 \setminus O_i^1$
Проверка условия $C_k^1 \cong C_i^1$ (или $C_k^0 \cong C_i^0$) для некоторого $k > i$
При выполнении − стоп
Искомая модель есть модель M_k. При этом класс $K_0 = C_{k-1}^0$, а класс $K_1 = C_{k-1}^1$ − результирующее множество «забракованных» объектов
При выполнении условия $K_1/m > 3$ запускается аналогичный итерационный процесс для множества K_1 и т. д.

[1] Поскольку, как сказано в работе [1], в силу несостоятельности коэффициентов модели мы не можем пользоваться понятием репрезентативности выборки, под «достаточной мощностью» класса K_1 будем (для конкретности) понимать выполнение условия $(|K_1|:m) > 3$ (см. табл. 1). Оно выполняется, в частности, в приводимом далее примере.

на множество K_{11} объектов, ОПО которых моделью $M_{1нач}$ не превышает δ_{max}, и множество K_{12} «забракованных» объектов и запустить аналогичный итерационный процесс. Результатом этого процесса будет результирующая модель $M_{1рез}$ с требуемым качеством оценки и класс K_2 объектов, «забракованных» этой моделью. Далее аналогичным образом рассматриваем класс K_2 и т. д. Здесь полезно посмотреть начало приводимого далее примера (см. § 5).

Ключевую роль в предлагаемом многомодельном подходе к массовой оценке играет процедура классификации оцениваемых объектов. Чтобы описанный в работе [1] эвристический алгоритм классификации был работоспособен для большего числа классов, в него внесены существенные изменения, рассматриваемые далее. Теоретическое обоснование работы алгоритма требует отдельного рассмотрения.

2. МОДИФИКАЦИЯ АЛГОРИТМА КЛАССИФИКАЦИИ ДЛЯ СЛУЧАЯ БОЛЕЕ ДВУХ КЛАССОВ

Далее кратко описываются основные этапы модифицированного эвристического алгоритма классификации и определяются его основные характеристики. Для определения и настройки количественных показателей, характеризующих качество алгоритма, вначале рассматривается процедура классификации объектов РВ (т. е. объектов с известной принадлежностью к тому или иному классу), а затем — применение настроенного алгоритма к оцениваемым объектам. Некоторые этапы — те же, что и в алгоритме из работы [1], а некоторые — либо дополнены, либо полностью изменены. Напомним ключевое используемое понятие d-близости. Значение a_1 некоторого количественного фактора a называется *d-близким* ($d > 0$) к значению a_2, если справедливо $|a_2 - a_1| \le da_2$. Отношение d-близости в общем случае несимметрично. При ограниченной длине выборки РВ число d-близких к a_2 значений фактора a пропорционально значению кумулятивной вероятностной функции для a_2 в точке, удаленной от a_2 на расстояние d. Далее приводятся основные этапы предлагаемого алгоритма классификации.

А. Использование понятия d-близости при рассмотрении значений факторов стоимости объектов РВ. Каждый из объектов РВ выбирается независимо, и с его описанием сравниваются описания остальных объектов РВ. Обозначим через ВО очередной выбираемый объект. Описание ВО в разрезе значений факторов стоимости сравнивается с описанием каждого из объектов РВ. При этом последовательно проходят этапы:

1) для значения x_{ij} фактора X_i, $1 \le i \le m$, из описания ВО находится совокупность $S(x_{ij}, d)$ объектов из РВ с d-близкими к x_{ij} значениям фактора X_i; показатель d-близости выбирается, начиная с малым значением Δ, с шагом Δ ($d(k) = k\Delta$) и возрастает до тех пор, пока число объектов РВ в совокупности $S(x_{ij}, d)$ не превзойдет задаваемого числа[2] G, достигнутое значение d фиксируется и обозначается как $d_{ij} = g_{ij}\Delta$, где g_{ij} отлично от нуля при выполнения условия $\min|S(x_{ij}, d)| > G$.

2) По совокупности $S(x_{ij}, d_{ij})$ определяются показатели: $k_\nu(x_{ij}, d_{ij})$, $\nu = 1, ..., N$, где $k_\nu(x_{ij}, d_{ij})$ — число d_{ij}-близких к x_{ij} значений фактора X_i, входящих в описания объектов из класса[3] K_ν, а N — число классов;

3) эти показатели нормируются на соответствующие количества объектов $|K_\nu|$ в каждом из классов K_ν, в результате получаем величины $p_\nu(x_{ij}, d_{ij})$, пропорциональные частотам вхождения значений фактора X_i, d_{ij}-близких к значению x_{ij}, в классы K_ν;

4) используем эвристическое соображение: «чем ближе между собой значения факторов стоимости, принадлежащие разным объектам, тем чаще такие объекты принадлежат одному и тому же классу». Поэтому взвешиваем каждую величину $p_\nu(x_{ij}, d_{ij})$ с весом $1/\log(g_{ij})$

Б. Построение информационной матрицы. Для каждого фактора X_i, $i = 1, ..., m$, найденные N показателей вида $p_\nu(x_{ij}, d_{ij})/\log(g_{ij},)$, $\nu = 1, ..., N$, образуют i-й столбец *информационной матрицы* (ИМ). Таким образом, каждый ВО из РВ характеризуется своей ИМ, строки которой соответствуют классам K_ν, а столбцы — m факторам стоимости, используемым в моделях, построенных в § 1.

В. Построение классифицирующей матрицы. Найдем суммы элементов по каждой из строк ИМ, соответствующий N-вектор указывает на «шансы» принадлежности данного объекта к тому или иному классу. Сформируем *классифицирующую мат-*

[2] Выбор значения G особых трудностей не представляет и требует нескольких прогонов алгоритма с разными значениями G, при которых точность классификации объектов из РВ (как функция от G) имеет экстремум (или «плато» экстремальных точек). Искомое значение G выбирается как одна из таких экстремальных точек.

[3] Поскольку в алгоритме классификации каждая строка информационной матрицы (см. далее) однозначно соответствует номеру класса, исходный класс K_0 «пригодных» объектов будет далее (для конкретности) обозначаться как K_2.

рицу (КМ), строки которой соответствуют объектам РВ, причем каждая строка в КМ является N-вектором, полученным по соответствующей данному объекту ИМ, так что КМ имеет размер $(n \times N)$, где n — длина РВ. Сопоставим КМ линейную регрессионную модель $M_{\text{КМ}}$, в которой каждый столбец КМ рассматривается как соответствующий предиктор для модели $M_{\text{КМ}}$. В качестве зависимой переменной в модели $M_{\text{КМ}}$ возьмем n-вектор, координаты которого суть номера классов, полученных в § 1 на РВ в результате применения итерационной процедуры построения моделей. Прежде чем анализировать модель $M_{\text{КМ}}$, рассмотрим, как анализировались подобные модели в известных публикациях.

Традиционный описанный в литературе подход к анализу моделей с категориальным выходом состоит в следующем. Поскольку в таких моделях истинное (непрерывное) значение зависимой переменной не наблюдаемо в пределах каждой категории, ненаблюдаемой переменной сопоставляется то или иное вероятностное распределение, характеризующее вероятность попадания этой переменной в ту или иную категорию, что обусловливает соответствующее наблюдаемое значение y_i модели. Наиболее часто в литературе используются кумулятивные вероятностные распределения, удовлетворяющие условию [2] $Pr[y \le \nu | X] = F(b_\nu - X^T\beta)$, $\nu = 1, \ldots, N-1$, где b_ν и β — неизвестные параметры модели, а F — некоторая монотонно возрастающая функция, отображающая действительную прямую в единичный интервал. В классе таких функций наибольшее распространение получили стандартные нормальные распределения и логистические распределения — соответствующие модели называются ординальными (упорядоченными) пробит- и логит-моделями. Появились также многочисленные вариации этих моделей (см., например, работу [3]).

В принципе, можно пойти по традиционному пути и в роли классификатора использовать какую-нибудь соответствующую, скажем, логистическую модель — либо вообще вместо алгоритма классификации, либо только для КМ. Весь вопрос — в точности классификации (определение см. далее).

Далее описывается альтернативный подход к анализу качества модели $M_{\text{КМ}}$ как классификатора для КМ, который (как показано в § 5) может обеспечить большую точность классификации по сравнению с соответствующей логистической моделью.

3. АЛЬТЕРНАТИВНЫЙ ПОДХОД К ИСПОЛЬЗОВАНИЮ РЕГРЕССИОННОЙ МОДЕЛИ $M_{\text{КМ}}$

Дополним КМ столбцом из единиц и будем рассматривать расширенную КМ (обозначаемую как $M_{\text{РВ}}$) как исходный массив для построения линейной регрессионной модели $M_{\text{КМ}}$ (со свободным членом). Наблюдаемой зависимой переменной Z модели $M_{\text{КМ}}$ является номер класса (категории), определенный на объектах РВ. Напомним, что каждый столбец в КМ, рассматриваемый как предиктор модели $M_{\text{КМ}}$, есть n-вектор с непрерывными действительными координатами. В предположении, что номера классов суть действительные числа из R, с помощью обыкновенного метода наименьших квадратов найдем вектор коэффициентов $\beta_Z = (M_{\text{РВ}}^T M_{\text{РВ}})^{-1} M_{\text{РВ}}^T Z$ этой модели. Обозначим через $Z_{\text{непр}} = M_{\text{РВ}}\beta_Z$ совокупность вычисленных значений, несущих скрытую информацию о значениях зависимой переменной Z модели $M_{\text{КМ}}$ и принадлежащих некоторому интервалу $[0, R_{\text{КМ}}] \subset R$, содержащему числа 1, ..., N. Как и в традиционном подходе, в силу категориальности переменной Z значения коэффициентов β_Z, а, следовательно, и функции $Z_{\text{непр}}$ содержат неизвестные ошибки.

Далее нам потребуется важное свойство регрессионной модели, называемое в литературе *инвариантностью к перестановкам* (*exchangeability* [4]). Применительно к модели $M_{\text{КМ}}$ оно означает выполнение условий:

$$Z(p(i)) = p(Z(i)),$$

$$Z_{\text{непр}}(p(i)) = p(Z_{\text{непр}}(i)),$$

где i — номер объекта в РВ, а p — произвольная перестановка на множестве объектов РВ.

В нашем случае задача состоит в том, чтобы значения функции $Z_{\text{непр}}(i)$ округлить до целых чисел 1, 2, ..., N некоторым способом, обеспечивающим наибольшее совпадение функции $\hat{Z}_{\text{непр}}$ (с округленными значениями) с функцией Z. Алгоритм требуемого округления значений функции $Z_{\text{непр}}(i)$:

1) значения $Z_{\text{непр}}(i)$ упорядочиваются по возрастанию;

2) для каждого интервала[4] $[\nu, \nu + 1]$, $\nu = 1, \ldots, N-1$, находится полная совокупность K_ν объектов

[4] В силу инвариантности по перестановкам каждый из интервалов $[\nu, \nu + 1]$ может рассматриваться независимо от остальных интервалов.

i таких, что $Z_{\text{непр}}(p(i)) \in [\nu, \nu + 1]$ (где p — перестановка объектов из РВ, связанная с п. 1);

3) в каждом интервале $[\nu, \nu + 1]$, $\nu = 1, ..., N - 1$, определяется значение b_ν функции $Z_{\text{непр}}(p(i))$ такое, что:

а) все значения функции $Z_{\text{непр}}(p(i))$, меньшие или равные (большие) b_ν, округляются до ν (до $\nu + 1$);

б) число совпадений округленных таким образом значений функции $\hat{Z}_{\text{непр}}(p(i))$ с соответствующими значениями функции $Z(p(i))$ — наибольшее для выбираемого значения b_ν;

4) сумма таких наибольших совпадений по всем интервалам $[\nu, \nu + 1]$, $\nu = 1, ..., N - 1$, характеризует «интегральное» качество алгоритма классификации.

Основным параметром алгоритма классификации является точность классификации, определяемая путем «прогонки» объектов с известной принадлежностью к классам (в данном случае — объектов РВ) через алгоритм классификации и сравнения полученного распределения классов на классифицированных объектах с исходным распределением. Точность классификации определяется как число правильно классифицированных объектов к их общему числу. Однако точность классификации служит лишь одним из качеств алгоритма классификации. При наличии более двух классов с сильно различающейся мощностью основную «нагрузку» при классификации может брать на себя класс с наибольшей мощностью, и высокий процент классификации может не означать хорошей классификации в классах с небольшой мощностью. Подобные ситуации целесообразно учитывать с использованием так называемого *коэффициента Криппендорфа* (см., например, работу [5], в которой рассмотрены процедуры вычисления коэффициента Криппендорфа для сравниваемых множеств различной природы: множеств бинарных элементов, номинальных и порядковых элементов, множеств с введенной метрикой и т. д.). В нашем случае используется процедура сравнения двух множеств с номинальными элементами (номерами классов). Подробности см. в § 5.

4. КЛАССИФИКАЦИЯ ОЦЕНИВАЕМЫХ ОБЪЕКТОВ И ИХ ОЦЕНКА СООТВЕТСТВУЮЩИМИ МОДЕЛЯМИ

Результатом анализа алгоритма классификации служат найденные численные значения некоторых параметров, используемые при классификации оцениваемых объектов:

— вектор β_Z коэффициентов модели $M_{\text{КМ}}$;

— разделяющие значения типа b_ν внутри каждого из интервалов $[\nu, \nu + 1]$, $\nu = 1, ..., N - 1$.

Обозначим через $M_{\text{ОО}}$ массив объектов, требующих оценки. Процедура оценки этих объектов и верификация результатов оценки состоит из следующих этапов (напомним, что факторы стоимости у объектов из $M_{\text{ОО}}$ те же, что и у построенных на РВ моделей).

A. «Прогон» каждого оцениваемого объекта из $M_{\text{ОО}}$ через алгоритм классификации с целью построения для него соответствующей ИМ, далее — построение для массива $M_{\text{ОО}}$ классифицирующей матрицы $\text{КМ}_{\text{ОО}}$ и по $\text{КМ}_{\text{ОО}}$ — матрицы типа матрицы $M_{\text{РВ}}$ с тем же числом $(m + 1)$ столбцов, что и у матрицы $M_{\text{РВ}}$.

B. С использованием найденного ранее вектора коэффициентов β_Z нахождение непрерывной функции (j), $j \in M_{\text{ОО}}$.

C. Учитывая найденные ранее разделяющие значения типа b_ν внутри каждого из интервалов $[\nu, \nu + 1]$, $\nu = 1, ..., N - 1$, округление непрерывных значений функции (j), $j \in M_{\text{ОО}}$ до целочисленных значений и отнесение объекта j к классу $\hat{Z}^{\text{ОО}}_{\text{непр}}(j)$.

D. Оценка объекта j с помощью модели, соответствующей классу $\hat{Z}^{\text{ОО}}_{\text{непр}}(j)$. В результате этой процедуры каждый объект из $M_{\text{ОО}}$ будет отнесен к какому-либо из N классов. Обозначим через $K^{\text{ОО}}_\nu$ множество объектов из $M_{\text{ОО}}$, отнесенных к классу с номером ν.

E. Верификация полученных классов на массиве $M_{\text{ОО}}$ производится в рамках дисперсионного анализа и линейного дискриминантного анализа Фишера (см., например, работу [6]). В роли дискриминантной функции для данного класса используется правая часть уравнения модели для этого класса. Требуемая верификация осуществляется следующим образом:

— для каждого класса $K^{\text{ОО}}_\nu$, $\nu = 1, ..., N$, вычисляются стоимости всех объектов из $K^{\text{ОО}}_\nu$ с помощью модели M_ν, соответствующей классу K_ν на РВ;

— вычисляются средние значения $m^{\text{ОО}}_\nu$ стоимостей и дисперсии $\sigma^{\text{ОО}}_\nu$ стоимостей для каждого $K^{\text{ОО}}_\nu$, аналогично вычисляются средние значения m_ν стоимостей и дисперсии σ_ν стоимостей для каждого класса K_ν из РВ;

— в рамках дисперсионного анализа сравниваются соответствующие пары (m_ν, m_ν^{OO}) и (σ_ν, σ_ν^{OO}) и определяется (не)значимость различия элементов в каждой паре[5] (т. е. (не)значимость соответствия $K_\nu \to K_\nu^{OO}$, $\nu = 1, ..., N$).

Практическое применение данной процедуры описано в следующем параграфе.

ПРИМЕР: МАССОВАЯ ОЦЕНКА УДЕЛЬНЫХ СТОИМОСТЕЙ КВАРТИР В РАЗНЫХ РАЙОНАХ г. СОЧИ

Исходные данные. Исходная выборка с рыночной информацией (РВ) содержала 101 объект (квартиры в многоквартирных домах). В качестве факторов стоимости моделей оценки были выбраны: 1 — площадь объекта; 2 — район местоположения объекта; 3 — расстояние до делового центра; 4 —расстояние до ж/д станции; 5 — расстояние до положительного центра притяжения; 6 — расстояние до берега моря. В качестве зависимой переменной рассматривалась стоимость 1 кв. м квартиры. Критерием качества оценки выбрана *относительная погрешность оценки* (ОПО), Начальное максимальное допустимое значение ОПО равнялось 15 %. На исходной РВ была построена линейная модель оценки, которая разделила все объекты РВ на два класса: класс C_0 «пригодных» объектов (с ОПО ≤ 15 %), содержащий 37 объектов, и класс C_1 «забракованных» объектов (с ОПО > 15 %), содержащий 64 объекта.

Итерационная процедура построения моделей оценки. Классы C_0 и C_1 рассматривались как исходное состояние итерационного процесса построения моделей, представленного в табл. 1. Число элементов в классе «пригодных» объектов на каждом шаге итерации показано на рис. 1. Видим, что мощность класса «пригодных» объектов возросла с 37 до 51 объекта, после чего по «пригодным» объектам процедура зациклилась на десятой итерации, так что результирующий класс K_0 «пригодных» объектов содержит 51 объект. Качество оценки исходной модели M_0^0 на объектах класса C_0^0 и результирующей модели M_{10}^0 на «пригодных» объектах из $K_0 = C_{10}^0$ показано соответственно в первой и второй строках табл. 2. С этого момента начинается аналогичная процедура поиска модели

[5] Эффективность такой процедуры зависит от точности алгоритма классификации и от степени расхождения описаний объектов из РВ и массива M_{OO} (от «биения» выборки [7]).

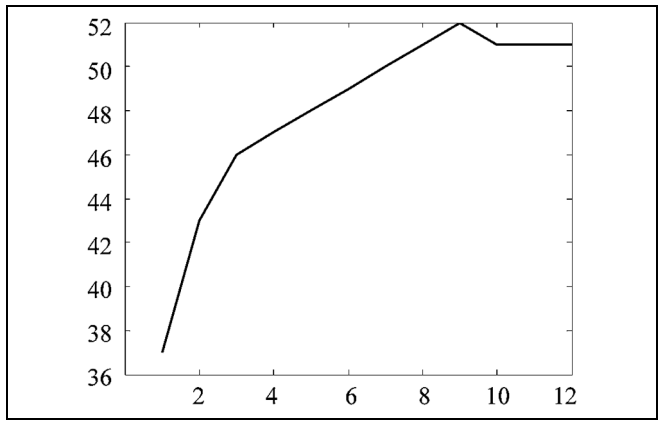

Рис. 1. Мощность класса «пригодных» объектов (с нулевыми верхними индексами)на каждом шаге итерации: ось абсцисс — номер итерации; ось ординат — мощность класса «пригодных» объектов)

для забракованных» объектов, начиная с множества C_{10}^1.

Однако при сохранении прежнего требования к ОПО (не более 15 %) классы «забракованных» объектов, пригодных для второй модели, получаются довольно мелкими, что затрудняет дальнейшее построение моделей. Поэтому для«забракованных» объектов, начиная с множества C_{10}^1, требования к качеству оценки были ослаблены: допустимая ОПО поднялась до 25 %. На следующем шаге итерации мощность класса «забракованных» объектов уменьшилась с 50 до 30 объектов, после чего процедура зациклилась. Таким образом, класс $K_1 = C_{11}^1$ ранее «забракованных» объектов, но «пригодных»для второй модели, содержит 20 объектов. Качество оценки результирующей модели M_{11}^1 на «пригодных» объектах из C_{11}^1 показано в третьей строке табл. 2.

«Забракованные» относительно модели M_{11}^1 объекты образуют класс K_3 (см. сноску 3), содер-

Таблица 2

Качество оценки моделей на «пригодных» и «забракованных» объектах

Показатели качества оценки	$\delta_{ср}$, %	R^2
Исходная модель M_0^0 на C_0^0	5,7	0,8590
Модель M_{10}^0 на C_{10}^0	6,8	0,7703
Модель M_1^1 на C_{11}^1	20,03	0,7089

жащий 30 объектов. Поскольку регрессионное оценивание объектов из K_3 характеризуется неприемлемо высокими значениями ОПО, для оценки этих объектов следует применять альтернативные методы.

Определение классифицирующей способности алгоритма. Согласно сказанному в § 2, все объекты из РВ были пропущены через алгоритм классификации, последовательность операций в котором можно представить в виде схемы:

$$\text{Объекты из РВ} \to \{\text{ИМ}(i)\} \to \text{КМ} \to M_{\text{РВ}} \to \\ \to \beta_Z \to Z_{\text{непр}}.$$

Результирующие данные алгоритма при $G = 6$: вектор β_Z коэффициентов модели $M_{\text{КМ}}$ β_Z = (2,1202; −1,4171; −0,4641; 2,0927), разделяющие значения функции $Z_{\text{непр}}$ при округлении до целых чисел: $b_1 = 1{,}5102$, $b_2 = 2{,}3534$. Таким образом, в предположении упорядоченности категорий их границы суть: категория 1 — (0; 1,5102], категория 2 — (1,5102; 2,3534], категория 3 — (2,3534; ∞). Классифицирующая способность алгоритма определяется по данным, представленным в табл. 3.

Здесь K_i — исходные классы, определенные на РВ в процессе итерационного построения моделей, а K_i' — результирующие классы на выходе алгоритма классификации. Основные характеристики алгоритма классификации: точность = 100 % × × (8 + 43 + 23)/101 = 73,3 %. Коэффициент

Таблица 3

Классифицирующая способность алгоритма

Классы	$K_1' = 11$	$K_2' = 61$	$K_3' = 29$
$K_1 = 20$	8	11	1
$K_2 = 51$	3	43	5
$K_3 = 30$	0	7	23

Таблица 4

Средние значения и дисперсии стоимостей в классах K_1 и K_2 на РВ и в классах $K_1^{\text{ОО}}$ и $K_2^{\text{ОО}}$, определенных на массиве $M_{\text{ОО}}$

Массив	Среднее значение стоимостей в классах, руб.		Дисперсия стоимостей в классах	
РВ	118090	81469	$3{,}6038 \cdot 10^7$	$8{,}1867 \cdot 10^7$
$M_{\text{ОО}}$	99233	72313	$3{,}0636 \cdot 10^7$	$10{,}898 \cdot 10^7$

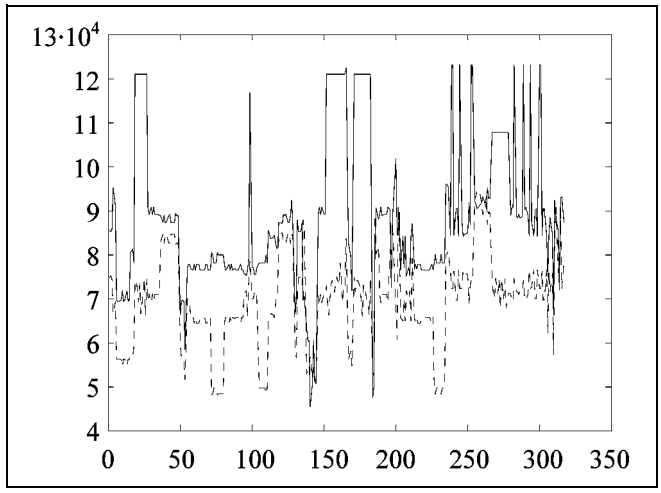

Рис. 2. Оценка объектов из классов $K_1^{\text{ОО}}$ и $K_2^{\text{ОО}}$ соответствующими моделями: сплошная линия — с использованием двух моделей M_{10}^0 и M_{11}^1 ;штриховая линия — с использованием одной модели M_0^0 , ось абсцисс — объекты из $K_1^{\text{ОО}}$ и $K_2^{\text{ОО}}$; ось ординат — модельная удельная стоимость

Криппендорфа[6] [5] равен 0,5443, что в качественной шкале для этого коэффициента можно трактовать как «хорошее» качество классификации.

Для сравнения классификационной способности предложенного алгоритма с традиционным подходом к построению классификатора с помощью логистической модели по КМ была построена логистическая модель *with proportionalodds*[7] (см., например, работу [8]) и с той же зависимой переменной Z. Опять же в предположении упорядоченности классов) были получены значения b_1 и b_2 для их границ, а именно: $b_1 = 0{,}4902$, $b_2 = 2{,}7506$, так что упорядоченные категории выглядят следующим образом: категория 1 — (0; 0,4902], категория 2 — (0,4902; 2,7506], категория 3 — (2,7506; ∞). Точность классификации логистической модели при этом равна 57 % (меньше 73,3 %).

Классификация объектов массива $M_{\text{ОО}}$. Массив $M_{\text{ОО}}$ оцениваемых объектов представляет выборку из 401 объекта с теми же факторами стоимости, что и в построенных моделях оценки. Массив $M_{\text{ОО}}$ аналогичным предыдущему образом пропускается через алгоритм классификации при полученных для РВ настройках (G, β_Z, b_1, b_2) алгоритма клас-

[6] Ограниченный объем статьи не позволяет изложить методику вычисления коэффициентов Криппендорфа.

[7] В отечественной литературе нет устоявшегося термина для этого типа моделей.

сификации. Результаты классификации: к классу K_1 отнесено 62, к классу K_2 — 255 и к классу K_3 — 84 объекта.

Оценивание объектов массива M_{OO}. Обозначим через β_Z^ν вектор коэффициентов модели, с помощью которой оцениваются объекты из массива M_{OO}, относящиеся к классу K_ν^{OO}, $\nu = 1, 2$: $\beta_Z^1 = 10^5(1{,}3701; 0{,}0000; -0{,}0212; -0{,}0000; -0{,}0002; -0{,}0000; -0{,}0000)$, $\beta_Z^2 = 10^4(9{,}8422; 0{,}0154; -0{,}5089; -0{,}0001; -0{,}0005; -0{,}0000; -0{,}0005)$. Объекты класса K_3^{OO} выделены в отдельное множество ввиду крайне плохого их оценивания регрессионными моделями. Графики удельных стоимостей объектов из классов K_1^{OO} и K_2^{OO} представлены на рис. 2.

Установление значимости соответствий $K_\nu \to K_\nu^{OO}$, $\nu = 1, 2$. Средние значения стоимостей и дисперсии стоимостей объектов из классов K_1 и K_2 на РВ и классов K_1^{OO} и K_2^{OO} на массиве M_{OO} приведены в табл. 4. Видно, что дисперсии между соответствующими классами K_1 и K_1^{OO}, а также между классами K_2 и K_2^{OO} разнятся существенно меньше, чем между классами K_1 и K_2^{OO}, а также между классами K_2 и K_1^{OO}. Это говорит о том, что оценивание объектов из массива M_{OO} с помощью двух моделей M_{10}^0 и M_{11}^1 более правдоподобно, нежели с помощью одной модели M_0^0.

ЗАКЛЮЧЕНИЕ

Предложенный многомодельный подход к массовой оценке позволяет получать более правдоподобные результаты при оценке объектов в больших массивах. Ключевую роль при этом играет алгоритм классификации, с помощью которого объекты массива разбиваются на классы, для каждого их которых используется соответствующая модель оценки, построенная по исходной выборке с рыночной информацией. Очень важный момент, привнесенный в массовую оценку в рамках многомодельного подхода, заключается в верификации результатов оценки объектов массива с помощью процедур дисперсионного и дискриминантного анализа. Возможность такой верификации отличает многомодельный подход от известных подходов к массовой оценке.

ЛИТЕРАТУРА

1. *Корноушенко Е.К.* Регрессионный подход к массовой оценке при наличии ненаблюдаемой составляющей в зависимой переменной // Проблемы управления. — 2013. — № 4. — С. 23—31.
2. *Boes S., Winkelmann R.* Ordered Response Models // Working Paper N 0507, Socioeconomic Institute, University of Zurich, 2005. — URL: www.soi.uzh.ch/research/wp/2005/wp0507.pdf (дата обращения 10.03.2013).
3. *Anans C.V., Kleinbaum D.G.* Regression Models for Ordinal Responses: A Review of Methods and Applications // Intern. J. Epidemology. — 1997. — Vol. 26, N 6. — P. 1323—1333. — URL: www.biostat.sdu.dk/.../misc/ordinalResponseModels.pdf (дата обращения 10.09.2013).
4. *McCullagh P.* Exchangeability and regression models // Univ. Chicago. Dept. Statistics, Techn. Report, N 544, 2004. — URL: www.stat.uchicago.edu/~pmcc/reports/exchangeability.pdf (дата обращения 18.02.2013).
5. *Krippendorff K.* Computing Krippendorff's Alpha-Reliability. — URL: www.asc.upenn.edu/usr/krippendorff/mwebreliability4.pdf (дата обращения 20.02.2013).
6. *Discover* Which Variables Discriminate Between Gtoups, Discriminant Function Analysis. — URL: www.statsoft.com/textbook/discriminant-function-analysis (дата обращения 10.03.2013).
7. *Huang J.*, et al. Correcting Sample Selection Bias by Unlabeled Data. — URL: www.books.nips.cc/papers/files/nips19/NIPS2006_0915.pdf (дата обращения 24.3.2013).
8. *Ordered* Logit Models — Overview. — URL: www3.nd.edu/~rwilliam/stats3/L11.pdf (дата обращения). 8.04.2013).

Статья представлена к публикации членом редколлегии Р.М. Нижегородцевым.

Корноушенко Евгений Константинович — д-р техн. наук, гл. науч. сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, г. Москва,
☎ (495) 334-90-00, ✉ekorno@mail.ru.