УДК 519.25

РЕГРЕССИОННЫЙ ПОДХОД К МАССОВОЙ ОЦЕНКЕ ПРИ НАЛИЧИИ НЕНАБЛЮДАЕМОЙ СОСТАВЛЯЮЩЕЙ В ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Е.К. Корноушенко

Предложен новый подход к массовой оценке в предположении, что зависимая переменная модели оценки содержит ненаблюдаемую составляющую, обусловленную, в частности, слабой развитостью рынка оцениваемых объектов. Подход включает в себя несколько важных этапов: построение двух моделей с требуемым качеством оценки, определенных на соответствующих непересекающихся множествах (классах); отнесение объектов, требующих оценки, к тому или иному из этих множеств (классификация); выбор соответствующей модели для оценки каждого из таких объектов. Показано, что данный подход позволяет полнее использовать рыночную информацию и существенно улучшить качество массовой оценки.

Ключевые слова: регрессионная модель, *d*-близость, классификация, точность, надежность, массовая оценка.

ВВЕДЕНИЕ

В практических целях регрессионное оценивание (чаще употребляется термин *массовая оценка*) применяется для оценки больших массивов объектов, когда индивидуальный подход к оценке каждого объекта нецелесообразен по экономическим соображениям. Для конкретности далее под массовой оценкой будем понимать массовую (кадастровую) оценку объектов недвижимости. Массовая оценка начинается со сбора рыночной информации на рассматриваемой территории. Рыночная информация должна содержать данные: о характеристиках объекта, его местоположении административной принадлежности и др., а также какую-либо информацию о стоимости объекта (например, цену предложения, цену сделки, цену аренды и т. п.). Далее, для краткости, будем пользоваться термином *стоимость*, (конкретный вид стоимости уточняется в каждом конкретном случае). Эта информации используется для формирования *рыночной выборки* (РВ), по которой строится регрессионная модель, применяемая далее для оценки объектов на рассматриваемой территории. В качестве зависимой переменной такой модели используется конкретный вид стоимости объектов из рыночной информации. В общем случае значения стоимости могут зависеть от некоторых не учитываемых при оценке факторов, от неразвитости рынка и пр., что обусловливает наличие ненаблюдаемой составляющей в рыночных значениях стоимости. Ошибка регрессионной модели оценивания будет состоять из двух слагаемых: собственной ошибки регрессионной модели и ошибки, обусловленной наличием ненаблюдаемой составляющей в зависимой переменной. Если коэффициенты строящейся модели вычисляются с помощью какого-либо из методов наименьших квадратов (линейного или нелинейного), наличие ненаблюдаемой составляющей в зависимой переменной приводит, согласно работе [1], к несостоятельности этих коэффициентов. Но тогда для построения регрессионной модели с хорошим качеством оценивания необходима «жесткая» фильтрация[1] РВ путем последовательных удалений из нее «забракованных» объектов, влияющих на качество модели (своеобразный аналог пошаговой регрессии с удалением объектов). При достижении

[1] Можно предположить, что такая фильтрация неявно означает, что из РВ удаляются, прежде всего, объекты с большими ненаблюдаемыми составляющими в зависимой переменной.

требуемого качества оценки в РВ остается определенное количество «забракованных» объектов, «не вошедших» в результирующую модель. Удаление объектов из РВ, особенно при ее небольших размерах, является негативным моментом, ухудшающим репрезентативность результирующей выборки, и в силу этого — ухудшение статистических свойств построенной модели.

Построение модели с хорошим качеством оценивания особенно важно при массовой оценке, когда построенная модель применяется для оценки больших массивов объектов. Если в процессе построения модели в РВ осталась некоторая доля «забракованных» объектов, не вошедших в модель, и если в РВ отражены основные свойства объектов оцениваемого далее массива[2], то справедливо предположить, что в общем случае примерно такая же доля объектов массива может быть отнесена к «забракованным» объектам, и оценка таких объектов используемой моделью будет характеризоваться большими ошибками оценивания.

При проведении массовой оценки за рубежом широко применяются электронные карты территории [2], позволяющие сделать привязку оцениваемых объектов к конкретному местоположению. Подобная привязка позволяет по рыночной информации строить так называемые оценочные зоны с тем или иным диапазоном рыночных цен и относить оцениваемый объект к той или иной оценочной зоне (после чего начинается учет остальных характеристик объекта). В России, к сожалению, в большинстве периферийных регионов с неразвитыми рынками массовая оценка земель и объектов недвижимости проводится либо без применения электронных карт, либо их применение только-только начинается.

Суть данной работы состоит в описании многомодельного (в частности, двухмодельного) подхода к массовой оценке, при котором объекты оцениваемого массива разбиваются на несколько классов (в частности, на два) и для каждого из этих классов по РВ строится своя модель оценки. Каждый из оцениваемых объектов относится вначале к тому или иному классу (с помощью предложенного далее алгоритма классификации), после чего этот объект оценивается соответствующей моделью. Такая дифференцированная оценка позволяет существенно улучшить качество массовой оценки, т. е. получить более точную информацию об оценках объектов массива по сравнению с довольно общими, как правило, замечаниями экспертов.

[2] Здесь имеется в виду, прежде всего, близость описаний объектов РВ и оцениваемого массива (подробнее см. § 3).

Для иллюстрации этого подхода приводится практический пример[3] массовой оценки промышленных объектов в сельских населенных пунктах Калужской области.

1. ИТЕРАЦИОННАЯ ПРОЦЕДУРА ПОСТРОЕНИЯ МОДЕЛЕЙ ДЛЯ МАССОВОЙ ОЦЕНКИ

Перед построением регрессионной модели для массовой оценки из РВ предварительно удаляются объекты с явными выбросами в значениях стоимости и в значениях факторов стоимости. Пусть оставшаяся часть РВ содержит n объектов. Для простоты изложения в качестве регрессионной модели для массовой оценки рассмотрим линейную модель с m, $m < n$, факторами стоимости:

$$Y = X\beta + \varepsilon, \quad (1)$$

$$Y^* = Y + \Delta Y. \quad (2)$$

Здесь Y — «истинная[4]» стоимость (зависимая переменная), X — $(n \times m + 1)$-матрица преобразованных значений[5] факторов стоимости объектов из РВ (содержащая столбец из единиц для учета свободного члена модели), β — $(m + 1)$-вектор коэффициентов модели, ε — n-вектор ошибок модели. В нашем случае положение осложняется тем, что вместо «истинных» значений Y для измерений доступна лишь сумма (2) с ненаблюдаемыми значениями Y и ΔY; вектор $\Delta Y = Y^* - Y$ называется [1] *ошибкой выборки*. Невязка модели (1), (2) $\varepsilon = Y^* - X\beta = Y - X\beta + \Delta Y$. Наличие ненаблюдаемой составляющей ΔY обусловливает гетероскедастичность модели (1), (2), при этом оценки коэффициентов β, находимые с помощью обычного метода наименьших квадратов, становятся несостоятельными [1]. Отсюда следует некорректность применения статистических критериев значимости модели и качества оценки[6]. В таких условиях

[3] Соответствующая рыночная информация была предоставлена автору организацией, проводившей массовую оценку объектов недвижимости.

[4] «Истинное» значение стоимости можно понимать как такое ее значение, к которому стремились бы значения стоимости объектов с одинаковым описанием при увеличении количества таких объектов на рассматриваемом рынке.

[5] Считаем, что при построении модели используются известные процедуры, улучшающие качество модели (см., например, работу [3]): нелинейное кодирование значений номинальных факторов (приписывание меток), обеспечение однонаправленности влияний количественных факторов на зависимую переменную и т. д.

[6] Показано, как можно оценить коэффициенты модели с помощью метода максимума правдоподобия в случаях, когда дисперсии ошибок выборки для объектов постоянны либо пропорциональны с некоторым неизвестным коэффициентом пропорциональности [1]. Оба этих предположения нереальны в условиях массовой оценки.

Таблица 1

Псевдокод итерационного процесса построения по РВ моделей с требуемым качеством оценки

Задание начальных условий: РВ, m, δ_{max}.

Построение модели M_0 на РВ: классы C_0^0, 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.

один из путей построения практической модели оценки заключается в последовательной фильтрации РВ с помощью итерационной процедуры, псевдокод которой показан в табл. 1. Здесь δ_{max} — предельно допустимое значение *относительной погрешности оценки* (ОПО), которое выбрано как критерий[7] качества оценки. По исходной РВ строится начальная модель M_0, которая разбивает РВ на два класса: класс C_0^0 объектов, ОПО которых моделью M_0 не превышает δ_{max}, и класс C_0^1 «забракованных» объектов. Затем запускается итерационный процесс, представленный в табл. 1. Заметим, что множество *наименований* факторов стоимости одно и то же для всех моделей на всех итерациях, тогда как множество *значений* факторов стоимости может изменяться от модели к модели.

В каждой итерации верхний нулевой индекс у множеств обозначает «пригодные» объекты (относительно модели, построенной на данной итерации), а единичный индекс — «забракованные». В силу конечности РВ этот 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. Поскольку число состояний такого автомата конечно, он в итоге «зациклится», т. е. его состояния (или состояние) будут повторяться. При этом в силу построения модели с верхними нулевыми индексами, построенные для одного и того же состояния такого автомата в разные моменты времени, будут совпадать. Аналогичное утверждение справедливо для моделей с верхними единичными индексами. (Заметим, что в приводимом далее примере такой автомат 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 на качество оценки, и класс K_1 «забракованных» объектов.

В принципе, для достаточно мощного[8] класса K_1 можно построить модель M_1, разбивающую его на множество K_{11} объектов, ОПО которых моделью M_1 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. Использование модели $M_{1рез}$ в дополнение к модели $M_{0рез}$ позволяет полнее учитывать информацию, содержащуюся в РВ, и уменьшить результирующее множество изначально «забракованных» в РВ объектов. Заметим также, что на любом шаге итерации процесс можно остановить при приемлемом для оценщика соотношении между множествами «пригодных» и «забракованных» объектов.

Замечание. При достаточно длинной РВ можно запустить итерационный процесс и для класса K_2, и т. д. Препятствуют такому продолжению два обстоятельства. Первое — экономическое: малые длины РВ, получаемые со слабо развитых российских периферийных рынков. Второе — методологическое: задачи классификации при наличии более двух классов намного сложнее, чем при наличии двух классов, их практическое решение потребует гораздо больших затрат, чем получаемая до-

[7] Дело в том, что ОПО тесно связана с отношением стоимостей, определяемым как отношение модельной стоимости объекта к его рыночной, и рядом других коэффициентов, базирующихся на отношении стоимостей и характеризующих качество оценки.

[8] Поскольку, как уже сказано, в силу несостоятельности коэффициентов модели мы не можем пользоваться понятием репрезентативности выборки, под «достаточной мощностью» класса K_1 будем (для конкретности) понимать выполнение условия $|K_1|:m > 3$. На этом условии базируется двухмодельный подход к массовой оценке. Оно выполняется, в частности, в приводимом далее примере.

полнительная информация об оценке. Поэтому в данной работе мы ограничимся рассмотрением двух классов — K_0 и K_1, а объекты класса K_2 удаляем из РВ, так что класс K_2 не участвует в классификации объектов оцениваемого массива $M_{\text{ОО}}$. В практических задачах результирующие «забракованные» объекты могут иметь какие-либо характерные признаки, которые использует оценщик при выделении и оценке таких объектов в массиве $M_{\text{ОО}}$. ♦

Для классификации объектов из массива $M_{\text{ОО}}$ в данной работе используются описания объектов в разрезе факторов стоимости, характеризующих модели $M_{0\text{рез}}$ и $M_{1\text{рез}}$. Разбиение объектов РВ на классы K_0 и K_1 рассматривается как эталонное, а отнесение каждого объекта из массива $M_{\text{ОО}}$ к какому-либо из этих классов производится с помощью описываемого далее алгоритма классификации.

2. ОСОБЕННОСТИ КЛАССИФИКАЦИИ ОБЪЕКТОВ РЫНОЧНОЙ ВЫБОРКИ

2.1. Характеристика классов K_0 и K_1 объектов рыночной выборки

Справедливо предположить, что случайный или неслучайный характер состава классов K_0 и K_1 зависит от размеров ненаблюдаемой составляющей в стоимости объектов из РВ и, конечно же, от описаний объектов в разрезе факторов стоимости. О возможности неслучайного характера состава классов K_0 и K_1 говорят результаты представленного далее практического примера. Предполагается, что в РВ всякое значение фактора стоимости может принадлежать объектам из класса K_0 и объектам из класса K_1. В такой ситуации можно считать, что принадлежность объекта к классу K_0 или K_1 определяется той или иной комбинацией значений факторов стоимости, причем каждый класс характеризуется некоторой совокупностью «допустимых» для него комбинаций. На выявление таких комбинаций и направлен описываемый далее алгоритм классификации.

Отметим, что необходимо:

1) все используемые в модели факторы стоимости представлять в количественном виде; подобные преобразования факторов стоимости (практикуемые оценщиками для улучшения качества строящейся модели — см. сноску 5) должны быть сделаны на этапе построения модели M_0;

2) чтобы все количественные значения факторов стоимости были положительными; к нулевым значениям двоичных факторов прибавлять некоторую положительную константу (скажем, 2).

Ключевым в алгоритме является понятие *d-близости* значений факторов. Значение a_1 некоторого количественного фактора a называется *d-близким* ($d > 0$) к значению a_2, если справедливо $|a_2 - a_1| \leqslant da_2$. 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 d.

2.2. Классификация объектов рыночной выборки с помощью предлагаемого алгоритма

Кратко опишем основные этапы алгоритма классификации и определим его основные характеристики.

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 преобразованных значений (см. в п. 2.1 условия 1 и 2) факторов стоимости сравнивается с описанием каждого из объектов РВ. Последовательно выполняются следующие этапы:

1) для значения x_{ij} фактора X_i, $1 \leqslant i \leqslant m$, из описания ВО находится совокупность $S(x_{ij}, d_{ij})$ объектов из РВ с d_{ij}-близкими к x_{ij} значениями фактора X_i; показатель d_{ij}-близости выбирается таким, чтобы число объектов РВ в совокупности $S(x_{ij}, d_{ij})$ было не меньше задаваемого числа G (о выборе значения G будет сказано далее);

2) по совокупности $S(x_{ij}, d_{ij})$ определяются показатели: $k_1(x_{ij}, d_{ij})$ — число d_{ij}-близких к x_{ij} значений фактора X_i, входящих в описания объектов из класса K_1, и $k_0(x_{ij}, d_{ij})$ — аналогичное число для объектов из класса K_0;

3) эти показатели нормируются на соответствующие количества объектов $|K_1|$ и $|K_0|$ в каждом из классов K_1 и K_0. В результате получаем величины $p_1(x_{ij}, d_{ij})$ и $p_0(x_{ij}, d_{ij})$, пропорциональные частотам вхождения значений фактора X_i, d_{ij}-близких к значению x_{ij}, в классы K_1 и K_0;

4) поскольку мощность множества $S(x_{ij}, d_{ij})$ есть монотонно возрастающая функция от значений d_{ij}, возрастающих с шагом Δd_{ij}, обозначим через $d_i(G)$ значение d_{ij} при первом выполнении условия $|S(x_{ij}, d_{ij})| > G$;

5) на интервале $D(i, G) = [0, d_i(G)]$ строим графики значений показателей $p_1(x_{ij}, d_{ij})$ и $p_0(x_{ij}, d_{ij})$,

которые пропорциональны значениям соответствующих кумулятивных вероятностных функций;

6) по этим графикам определяется значение $d_i^* = \max\limits_{d \in D(i, G)} |p_1(x_{ij}, d_{ij}) - p_0(x_{ij}, d_{ij})|$ и находятся значения $p_1(x_{ij}, d_i^*)$ и $p_0(x_{ij}, d_i^*)$.

Классифицирующая матрица. Для каждого фактора X_i, $i = 1,..., m$, пара показателей $p_1(x_{ij}, d_i^*)$ и $p_0(x_{ij}, d_i^*)$ образует i-й столбец так называемой *классифицирующей матрицы* (КМ). Таким образом, КМ, строки которой соответствуют классам K_1 и K_0, имеет размер $(2 \times m)$.

Критерий классификации (дискриминантная функция). Пусть $K_{\text{ВО}}$ — номер класса, к которому будет отнесен ВО. Дискриминантная функция имеет вид: $K_{\text{ВО}} = \max(\Sigma_1 Q, \Sigma_0)$, где Σ_1 и Σ_0 — суммы элементов соответствующих строк КМ, а Q — параметр настройки алгоритма. Значение Q выбирается из условия: отношение числа N_1 объектов, классифицируемых алгоритмом как объекты класса K_1, к числу N_0, определяемому аналогичным образом, должно быть близким (или совпадать) с отношением $|K_1|/|K_0|$. В практических задачах поиск приемлемого значения параметра Q не вызывает трудностей: поскольку точного совпадения с отношением $|K_1|/|K_0|$ не требуется, для поиска приемлемого значения Q достаточно нескольких пробных значений Q, приводящих к значениям отношения N_1/N_0, большим и меньшим значениям отношения $|K_1|/|K_0|$. Простота процедуры нахождения приемлемого значения Q обусловлена очевидной монотонностью изменения значений соответствующих кумулятивных вероятностных функций в зависимости от показателей d-близости при изменении значения Q.

Предварительная настройка алгоритма. Описываемый алгоритм весьма чувствителен к выбору значения G. Дело в том, что для выполнения условия $|S(x_{ij}, d_{ij})| \geqslant G$ при увеличении G монотонно увеличивается и наибольшее значение d в каждом из множеств $S(x_{ij}, d)$. Но при этом в каждый класс начинают попадать объекты другого (смежного) класса, становящиеся d-близкими к x_{ij} для больших[9] d, т. е. «избирательная способность» d-близости падает. Таким образом, зависимость классифицирующей способности алгоритма от G имеет максимум при некотором значении G^*, для поиска которого достаточно также нескольких пробных шагов.

[9] А при очень больших d_{ij} в один класс могут попасть все объекты РВ.

2.3. Особенности предлагаемого алгоритма классификации

Из огромного множества алгоритмов классификации выберем лишь те, которые концептуально близки к описанному алгоритму, и покажем, что данный алгоритм отличается от подобных ему известных алгоритмов следующими особенностями.

• Концептуально наиболее близким к данному алгоритму можно считать алгоритмы CAEP (Classification Algorithm with Emerging Patterns) [4] и JEP-Classifier [5], разработанные Г. Донгом и его коллегами и базирующиеся на понятии *мерцающих образов* (emerging patterns, *EPs*). «Мерцающий образ» в приложении к РВ — это такая комбинация значений факторов в описаниях объектов из РВ, частота вхождения которой в описания объектов из разных классов, заданных на РВ, существенно разная. В CAEP эти частоты вычисляются для каждого *EP*, затем в каждом классе агрегируются (суммируются) соответствующие частоты. В итоге для каждого класса определяется совокупность *EPs*, суммарная частота появления которых в данном классе больше, чем их появление в другом классе. Таким образом, в алгоритме CAEP каждый класс в итоге характеризуется соответствующей совокупностью *EPs*. Однако в нашем случае ситуация сложнее, чем в работе [4], потому что:

— описания классифицируемых объектов массива $M_{\text{ОО}}$ (объектов оценки, ОО), как правило, отличаются от описаний объектов РВ — именно этим и обусловлен переход к понятию d-близости. Для каждого значения фактора из описания ОО алгоритм выделяет *подмножество* объектов из РВ со значениями этого фактора, d-близкими к рассматриваемому значению. Разбиение этого подмножества на классы существенно зависит от рассматриваемого *значения* фактора. Это означает, что в данном случае определение *EP* условное $(EP|x_{ij})$, зависящее от рассматриваемого значения x_{ij} фактора X_i ОО (а не от выборки РВ, как в работе [4]). Поскольку для другого значения аналогичные разбиения будут другими, в контексте алгоритма EP-Classifier [5] можно сказать, что в роли *EPs* в предложенном алгоритме выступают значения факторов стоимости из описаний каждого ОО, которые можно назвать «*jumping EPs*»;

— при описании алгоритма CAEP отмечается, что нахождение совокупностей *EPs*, являющихся представителями классов, — довольно трудоемкая процедура. В предлагаемом алгоритме эта процедура заменена несложной процедурой нахождения d-близких значений и использованием максимальных разностей в значениях, являющихся представителями разных классов;

— применяемое в алгоритме правило классификации (идентификатор класса) аналогично при-

меняемому в алгоритме CAEP: наибольшая строчная сумма элементов в КМ аналогична наибольшей сумме вхождений *EPs* в тот или иной класс.

• Другая особенность алгоритма — его адаптивность: при классификации очередного ОО внутренние параметры алгоритма (порядок выбора факторов стоимости, значения d_{ij}, структуры КМ) «подстраиваются» под классифицируемый ОО.

• Еще одна особенность связана с количественными факторами. Известны «неприятности», которые доставляют непрерывные (количественные) факторы при построении классификационных деревьев [6] или подсчете *EPs* [4, 5]. Традиционный путь — к разбиению непрерывных диапазонов значений таких факторов на конечное число интервалов, т. е. переход к дискретизированным факторам, причем процесс дискретизации также влечет сопутствующие неприятности [7]. В предлагаемом алгоритме эти неприятности «обойдены» с введением понятия *d*-близости и рассмотрением для каждого фактора конечных подвыборок длины, не меньшей *G*.

2.4. Оценка надежности классификации

Обозначим через $\pi = (\pi_0, \pi_1)$ результирующее разбиение РВ на классы K_0 и K_1, причем блоками разбиения π служат либо мощности этих классов, либо состав классов (там, где такая двойственность не вызывает непонимания). После «пропускания» РВ через алгоритм классификации для ряда объектов поменяется номер класса, к которому они будут принадлежать согласно классификации, т. е. на РВ будет определено другое разбиение $\pi' = (\pi'_0, \pi'_1)$. Обозначим через π_{00} совокупность объектов РВ, для которых номера классов в разбиениях π и π' совпадают и равны нулю, аналогично определяются совокупности π_{11}, π_{10} и π_{01}. Тогда результат классификации объектов РВ можно представить в виде табл. 2.

В нашем случае ситуация предельно простая: два номинальных класса, конечное число классифицируемых объектов и признаков (факторов стоимости). В этом случае общий подход к определению вероятностей (не)правильной классификации (описанный в частности, в работе [8]) с помощью скользящего (парзеновского) окна и вероятностной меры (не)правильной классификации сводится к непосредственному подсчету соответствующих вероятностей и выборочных частот. В итоге вероятности правильной ($p_{\text{пр}}$) и неправильной ($p_{\text{н}}$) классификации определяются как

$$p_{\text{пр}} = (\pi_{00} + \pi_{11})/n_{\text{РВ}}, \quad p_{\text{н}} = (\pi_{01} + \pi_{10})/n_{\text{РВ}}, \qquad (3)$$

где $n_{\text{РВ}} = n - |K_2|$ — число классифицируемых объектов в РВ.

Рассмотрим проблему оценки надежности классификации под углом оценки надежности «согласия» двух классификаторов: первый классифицирует объекты РВ с помощью разбиения π, а второй — с помощью разбиения π'. Для корректности принятой постановки должны выполняться определенные условия, совпадающие по сути с аналогичными условиями, необходимыми при оценке так называемого *каппа-коэффициента* (см., например, работу [9]), и адаптируемые к нашему случаю:

— классифицируемые объекты не зависимы друг от друга;

— каждый объект классифицируется независимо;

— разбиения π и π' образуют полное множество (других разбиений на РВ нет).

«Согласие» классификаторов характеризуется вероятностью $p_{\text{пр}}$, а «несогласие» — вероятностью $p_{\text{н}}$. Тогда надежность классификации можно оценить (по аналогии с каппа-коэффициентом) с помощью коэффициента

$$R_{\text{РВ}} = (p_{\text{пр}} - p_{\text{н}})/(1 - p_{\text{н}}). \qquad (4)$$

В нашем случае справедливо $p_{\text{пр}} + p_{\text{н}} = 1$. Но тогда[10] при $p_{\text{н}} < 0{,}5$; $R_{\text{РВ}} = 2 - 1/p_{\text{пр}} > 0$.

Значения $R_{\text{РВ}}$, принадлежащие интервалу [0, 1], считаются допустимыми. Нулевое значение свидетельствует об отсутствии классифицирующей способности алгоритма по отношению к данной РВ. И чем ближе значение $R_{\text{РВ}}$ к единице, тем выше и классифицирующая способность и надежность алгоритма классификации.

Таблица 2

Представление результатов классификации объектов рыночной выборки

π	π'	
	π'_0	π'_1
π_0	π_{00}	π_{01}
π_1	π_{10}	π_{11}

[10] К. Гвет в работе [10] показал, что оценки вида (4) (где $p_{\text{пр}} = x$ и $p_{\text{н}} = y$ — некоторые переменные из интервала [0, 1]) могут быть использованы лишь при $y < 0{,}5$, в противном случае при «хороших» значениях x. эти оценки принимают необъяснимые малые значения. В нашем случае условие $p_{\text{н}} > 0{,}5$ свидетельствует о плохой классифицирующей способности алгоритма и служит сигналом к прекращению дальнейшего рассмотрения.

3. ПРИМЕНЕНИЕ АЛГОРИТМА КЛАССИФИКАЦИИ К ДИФФЕРЕНЦИРОВАННОЙ ОЦЕНКЕ КЛАССИФИЦИРУЕМЫХ ОБЪЕКТОВ

Рассмотрим теперь основной аспект данной работы — дифференцированную оценку объектов массива $M_{\text{ОО}}$. Суть дифференцированной оценки представлена кратко во Введении при описании целей настоящей работы. Согласно п. 2.4, на РВ определено разбиение $\pi = (\pi_0, \pi_1)$. Алгоритм классификации каждому ОО $\in M_{\text{ОО}}$, рассматриваемому независимо от остальных, приписывает номер класса 0 или 1, в итоге на множестве $M_{\text{ОО}}$ получаем разбиение $\tau = (\tau_0, \tau_1)$. Встает вопрос о качестве классификации объектов из $M_{\text{ОО}}$, т. е. о составе классов разбиения $\tau = (\tau_0, \tau_1)$. Положение осложняется тем, что на множестве $M_{\text{ОО}}$ нет исходных классов K_0 и K_1, с помощью которых можно было бы определить качество классификации. Повторная классификация ОО при менее чем 100 %-й точности классификации применяемого алгоритма лишь увеличит степень неуверенности в результатах такой классификации.

Известны работы, где регрессионная модель, построенная по обучающей выборке, применяется для оценки объектов на тестовой выборке, на которой условные вероятности зависимой переменной от значений предикторов модели отличны от аналогичных распределений на обучающей выборке. Подобное отличие приводит к эффекту, называемому *биением выборки* (sample selection bias) (см., например, работу [11], где предлагается один из вариантов борьбы с биением выборки). В нашем случае эффект биения выборки сказывается в том, что в силу отличия описания объектов из массива $M_{\text{ОО}}$ от описания объектов РВ условные вероятности того или другого класса от значений факторов стоимости могут быть различными для РВ и массива $M_{\text{ОО}}$. С учетом этого факта при классификации объектов из массива $M_{\text{ОО}}$ алгоритм следует подстраивать (в плане выбора значений Q и G) к массиву $M_{\text{ОО}}$. Условия, при которых классифицируются ОО, можно рассматривать как простейшие аналоги соответствующих условий из работы [11]:

— соответствующие диапазоны значений факторов стоимости объектов из РВ и $M_{\text{ОО}}$ «не сильно» различаются;

— значения Q (при прежнем значении G) выбираются с учетом выполнения условия

$$\frac{\tau_1}{|M_{\text{ОО}}|} \approx \frac{|\pi_1|}{|\text{РВ}|}. \qquad (5)$$

Таблица 3

Результаты классификации объектов оценки

τ	τ'	
	τ'_0	τ'_1
τ_0 τ_1	τ_{00} τ_{10}	τ_{01} τ_{11}

В качестве коррекции результатов классификации ОО, повышающей достоверность результатов, рассмотрим выборку $W = [\text{РВ}, M_{\text{ОО}}]$, на которой определены два разбиения π (на РВ) и τ (на $M_{\text{ОО}}$), образующие исходное разбиение $\chi = (\chi_0, \chi_1)$ на выборке W: $\chi_0 = \pi_0 \cup \tau_0$, $\chi_1 = \pi_1 \cup \tau_1$. Алгоритм классификации применяется к выборке W, при этом разбиение χ изменяется до разбиения χ', в итоге получаем таблицу для разбиения χ, аналогичную табл. 2. Из сравнения элементов, входящих в блоки разбиения τ с этими же элементами, входящих в блоки разбиения χ', получаем табл. 3, характеризующую качество классификации объектов массива $M_{\text{ОО}}$. Здесь τ' — совокупности элементов разбиения τ, входящие в тот или иной блок разбиения χ', полученного «пропусканием» выборки W через алгоритм классификации. Конечная цель данного подхода состоит в оценке надежности классификации, представленной табл. 2 (по аналогии с тем, как это делалось в п. 2.4). Табл. 3 является интегральной оценкой правильности отнесения каждого из ОО к соответствующему классу K_0 или K_1.

В приводимом далее практическом примере массовой оценки показаны основные этапы предлагаемого подхода к массовой оценке.

4. ПРИМЕР: МАССОВАЯ ОЦЕНКА ОБЪЕКТОВ ПРОМЫШЛЕННОГО НАЗНАЧЕНИЯ В СЕЛЬСКИХ НАСЕЛЕННЫХ ПУНКТАХ КАЛУЖСКОЙ ОБЛАСТИ

Исходной информацией, предоставленной автору, служит РВ из 197 указанных в заголовке примера объектов. В роли решающего показателя качества оценки использовалась ОПО с допустимыми предельными значениями $\delta \leqslant 12$ %.

Итерационный процесс построения моделей. На РВ была построена линейная регрессионная модель (см. сноску 5) M_0^0 с 12-ю факторами стоимости, которая разбила РВ на два класса: класс C_0 «пригодных» объектов, содержащий 119 объектов, и класс C_1 — «забракованных» объектов, содержащий 78 объектов (78:197 ≈ 40 % РВ). Затем для класса C_0 запускался итерационный процесс (см. табл. 1). На рис. 1 показан график изменения мощности класса «пригодных» объектов на каждой итерации.

Видим, что на 10-й итерации процесс пришел в устойчивое состояние с неизменными далее множествами

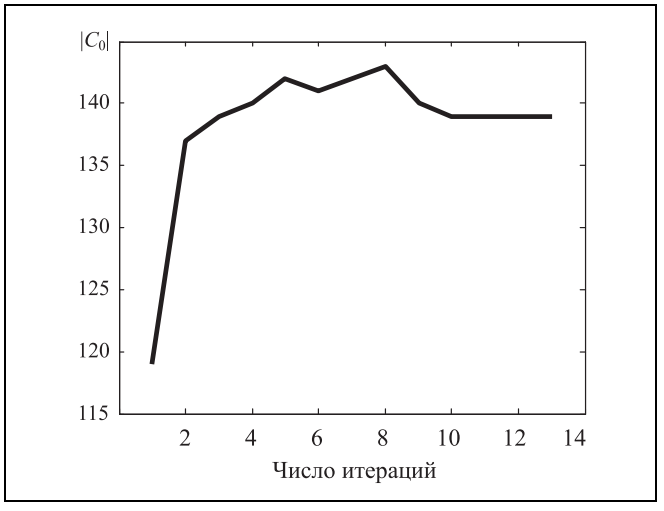

Рис. 1. Мощность класса «пригодных» объектов на каждой итерации

O_{10}^0 и O_{10}^1 соответственно «пригодных» и «забракованных» объектов. В этом состоянии для каждого из объектов множества O_{10}^0 из 139 объектов ОПО относительно модели на этом шаге итерации удовлетворяет требованию $\delta_i \leqslant 12$ %, и ни один объект к множеству O_9^0 не добавляется и из него не исключается. Теперь на множестве O_{10}^1 из 58 «забракованных» объектов строим регрессионную линейную модель M_0^1 и запускаем итерационный процесс для множества O_{10}^1. На первом же шаге итерации процесс пришел в устойчивое состояние: число «пригодных» относительно модели M_1^1 объектов равно 45, и далее это число не изменяется. При этом мощность класса K_2 «окончательно забракованных» объектов равна 13.

В табл. 4 приведены значения показателей качества (для наглядности приведены также соответствующие значения коэффициента детерминации R^2) для моделей M_1^0, M_{10}^0 и M_1^1. Поскольку модели M_{10}^0 и M_1^1 — результирующие для соответствующих итерационных циклов, это означает, что объекты РВ, не вошедшие в класс K_2, будут оцениваться соответствующими моделями с ОПО, не превышающей 12 %.

Настройка алгоритма классификации на РВ и определение надежности классификации. В результате предыдущего этапа на РВ определено разбиение π с блоками π_0 (139 объектов) и π_1 (45 объектов). При классификации объектов РВ значение Q выбирается из условия $\pi_1 \cong \pi_1'$, при этом $Q = 0{,}83$, а параметр $G = 10$. Результаты классификации объектов РВ с помощью алгоритма с указанными настройками приведены в табл. 5.

Согласно выражению (3) вероятность правильной классификации $p_{\text{пр}} = (116 + 20)/184 = 0{,}7391 > 0{,}5$, что вполне приемлемо для решаемой задачи, а надежность классификации согласно формуле (4) $R_{\text{РВ}} = 2/0{,}7391 = 0{,}6470$, что в качественной шкале надежности классификации считается «высокой».

Классификация объектов массива $M_{\text{ОО}}$. Применим теперь алгоритм к классификации объектов массива $M_{\text{ОО}}$, содержащего 100 ОО. В нашем случае при $Q_{\text{ОО}} = 0{,}79$ имеем: $\tau_1 = 25$, $\pi_1 = 45$, так что в соответствии с условием (5) $25/100 \approx 45/184 = 0{,}2446$. Для коррекции результатов классификации формируем выборку $W = [\text{РВ}, M_{\text{ОО}}]$, и после подстройки алгоритма классификации под выборку W ($Q_W = 0{,}85$) классифицируются объекты выборки W. Это позволяет извлечь необходимую информацию о вхождении каждого из объектов в тот или иной блок разбиения $\tau' = (\tau_0', \tau_1')$. В итоге получаем табл. 6, отражающую связь между разбиениями τ и τ'.

При этом коэффициент $R_{\text{ОО}}$, определяемый по аналогии с величиной $R_{\text{РВ}}$ и характеризующий надежность классификации, равен 0,9247, что можно считать высоким значением надежности классификации объектов. С учетом сказанного ранее о табл. 3 можно заключить, что (72 + 4) объекта «пригодные» для их оценки моделью M_{10}^1, а (3 + 21) объекта — моделью M_1^1. Оказалось, что в массиве $M_{\text{ОО}}$ присутствуют три объекта, для которых модель M_0^0 дала резкие выбросы в отрицательную

Таблица 4

Показатели качества оценки для разных моделей, построенных на «пригодных» объектах рыночной выборки

Показатели качества оценки	$\delta_{\text{ср}}$, %	R^2
Исходная модель M_1^0 на классе C_0^0	6,13	0,9440
Модель M_{10}^0 на классе C_{10}^0	5,91	0,9405
Модель M_1^1 на классе C_1^1	5,84	0,9109

Таблица 5

Результаты классификации объектов рыночной выборки

π	π'	
	141	43
139	116	23
45	25	20

Таблица 6

Результаты классификации объектов из массива $M_{\text{ОО}}$

τ	τ'	
	76	24
75	72	3
25	4	21

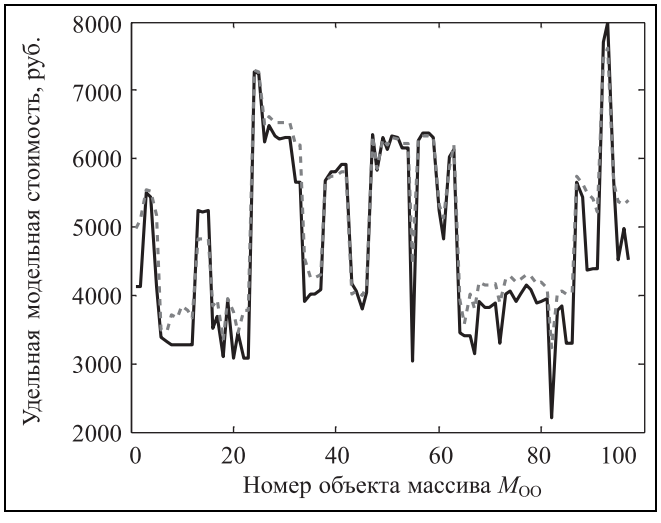

Рис. 2. Результаты оценки 97 объектов массива M_{OO}: сплошная — дифференцированная оценка ОО с помощью моделей M_{10}^1 и M_1^1, пунктир — оценка тех же объектов первоначальной моделью M_0^0

область. В то же время эти три объекта принадлежат множеству объектов, «пригодных» для модели M_1^1. Таким образом, с помощью моделей M_{10}^0 и M_1^1 оцениваются (с высокой надежностью) все объекты массива M_{OO}. На рис. 2 приведены результаты дифференцированной оценки 97 объектов с помощью моделей M_{10}^1 и M_1^1. Для сравнения приведены результаты оценки этих же объектов первоначальной моделью M_1^0.

Видим, что дифференцированная оценка объектов позволила скорректировать значения стоимостей многих ОО, полученных с помощью первоначальной модели M_0^0. Средняя относительная ошибка между этими графиками в данном случае составляет 8,4 %. Степень подобной коррекции существенно зависит от объектов массива M_{OO}.

ЗАКЛЮЧЕНИЕ

Предложен новый (двухмодельный) подход к массовой оценке объектов, позволяющий:

- с помощью итерационного процесса построения моделей полнее использовать информацию, содержащуюся в рыночной выборке;
- осуществлять дифференцированную оценку объектов, путем:

— построения отдельных моделей для «пригодных» и «забракованных» объектов рыночной выборки;

— классификации объектов на «пригодные» и «забракованные»;

— выбора соответствующей модели при оценке каждого объекта.

Дополнение существующей процедуры массовой оценки построением второй модели и этапом классификации объектов позволяет существенно улучшить точность массовой оценки. В настоящее время о точности массовой оценки зачастую судят по значениям некоторых интегральных показателей (типа среднего значения, максимальной или минимальной стоимости и т. п.), не опускаясь до оценки отдельных объектов.

ЛИТЕРАТУРА

1. *Lewis J.B., LinzerD.A.* Estimating Regression Models in Which the Dependent Variable Is Based on Estimates// Political Analysis. — 2005. — Vol.13. — P. 345—364. URL: http://www.sscnet.ucla.edu/polisci/faculty/lewis/#a_prereprint (дата обращения 04.06.2013).
2. *Ward R.D.,* et al. Improving CAMA Models Using Geographic Information Systems/Response Surface Analysis Location Factors// Assessment Journal. — 1999. — Vol. 31, N 1.
3. *Корноушенко Е.К.* Методологические аспекты практического регрессионного оценивания // Проблемы управления. — 2008. — № 2. — С. 34—41.
4. *Dong G.,* et al. CAEP: Classification by Aggregating Emerging Patterns // Discovery Sci. 99, LNAI 1721, Tokyo, Japan, 1999. URL: www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.3226 (дата обращения 07.06.2013).
5. *Li J.,* et al. Making Use of the Most Expressive Jumping Emerging Patterns for Classification // Proc.of Pacific Asia Conference on Knowledge Discovery in Databases (PAKDD), Kyoto, Japan, 2000. URL: www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.9640 (дата обращения 07.06.2013).
6. *Kohavi R., Quinlan J.R.* Improved Use of Continuous Attributes in C4.5 // Journal of Artifical Intelligence Research. — 1996. — N 4. — P. 77—90. www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.46.3240 (дата обращения 08.06.2013).
7. *Fayyad U.M., Irani K.B.* Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. — URL: www.yaroslavvb.com/papers/fayyad-discretization.pdf (дата обращения 04.06.2013).
8. *McDermott E. and Katagiri Sh.* A Parzen Window Based Derivation of Minimum Classification Error from the Theoretical Bayes Classification Risk. — URL: www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8450 (дата обращения 07.06.2013).
9. *Sim J., Wright C.C.* The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements // Phys. Ther. — Vol. 85. — P. 257—268. www.physther.org/content/85/3/257.full (дата обращения 04.06.2013).
10. *Gwet K.* Kappa Statistic is not Satisfactory for Assessing the Extent ofAgreement Between Raters // Statistical Methods For Inter-Rater Reliability Assessment. — April 2002. — N 1. URL: www.agreestat.com/.../kappa_statistic_is_not_satisfactory.pdf (дата обращения 07.06.2013).
11. *Huang J.,* et al. Correcting Sample Selection Bias by Unlabeled Data. — URL: www.enpub.fulton.asu.edu/cseml/07spring/Sample.pdf (дата обращения 07.06.2013).

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

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