УДК 303.732.4:519.237.5

А.Н. СОЛОШЕНКО

ОБОБЩЕНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ, ВЕСА КАТЕГОРИИ ПЕРЕМЕННОЙ И ИНДЕКСА ДЖИНИ ДЛЯ НЕПРЕРЫВНОЙ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ, ПРИНИМАЮЩЕЙ ВЕРОЯТНОСТНЫЕ ЗНАЧЕНИЯ

Аннотация. Рассмотрены оригинальные средства усовершенствования формул метода максимального правдоподобия для логистической регрессии, формулы веса категории переменной, формулы показателя значения информации и формулы индекса Джини для обеспечения возможности использования непрерывной целевой переменной, принимающей вероятностные значения. Методикой реализации исследования является использование непрерывных весовых функций с определенными ограничениями для подсчета обобщенного логарифма функции правдоподобия, его обобщенного вектора градиента и обобщенной матрицы Гессе, а также использование возможностей теории вероятностей для обобщения веса категории переменной и индекса Джини.

Ключевые слова: логистическая регрессия, вес категории переменной, индекс Джини, метод максимального правдоподобия, кредитный скоринг, анализ отклоненных заявок.

ВВЕДЕНИЕ

Одной из наиболее важных задач математического и статистического моделирования и прогнозирования является задача бинарной классификации входящих данных исходя из имеющейся информации об аналогичных данных с известными целевыми исходами, которые отвечают двум взаимно исключающим классам на примере индикатора реализации кредитного события: наступления или отсутствия дефолта в задачах кредитного скоринга [1]. Наиболее важной проблемой при скоринговом моделировании является учет и анализ отклоненных заявок (reject inference) — ранее входящих данных с неизвестным и ненаблюдаемым бинарным исходом, в целях обеспечения стабильности обучающей выборки относительно входящего потока информации в критериях распределений входящих параметров [2]. Для задач бинарной классификации в кредитном скоринге обычно используется бинарная логистическая регрессия [3] — частный случай категориальной логистической регрессии [4].

Недостатком классической бинарной логистической регрессии является постулирование и ограничение целевой переменной в области определения только бинарных значений [3–5]. Ограничение предполагает числа 0 и 1 единственными возможными значениями для фактической целевой переменной при обучении модели. На выходе модели получаем прогнозные значения целевой переменной как действительные числа — вероятности принадлежности к классу будущего единичного исхода. Область определения целевой переменной делает невозможным включение данных с неизвестным исходом в обучающую выборку как частично классифицированных данных с вероятностными прогнозными метками. Бинарный подход позволяет выполнять только взвешивание элементов выборки путем включения одного и того же элемента в обучающую выборку определенное число раз (одинаковое для всех элементов), возможно с разными исходами. Примером являются данные с неизвестными, но вероятностно прогнозируемыми исходами (соотношение включенных исходов для каждого элемента должно в какой-то степени соответствовать присвоенной вероятности прогноза). При использовании данного подхода выборку с известными исходами необходимо включать заданное число раз с фактическими бинарными исходами.

Поэтому актуальной является задача обобщения логистической регрессии для случая непрерывной целевой переменной, принимающей вероятностные значения, что в частном случае отвечает множеству бинарных значений (с вероятностью 0% и 100%), особенно для задач включения и анализа отклоненных заявок (reject inference) [6]. Также для обеспечения полного цикла построения скоринговой модели [5] актуальными являются вопросы определения показателя веса категории переменной и индекса Джини [5, 7] в терминах непрерывной целевой переменной, принимающей вероятностные значения. Актуальность рассматриваемой тематики затрагивает современные вопросы интеллектуального анализа данных и свидетельствует о необходимости расширения возможностей вероятностного вывода неизвестных значений целевой переменной [8].

ПОСТАНОВКА ЗАДАЧИ

Объектами исследования являются метод логистической регрессии (метод максимального правдоподобия), формула веса категории переменной и методы подсчета показателя Джини.

Предметом исследования является обобщение метода максимального правдоподобия относительно моделирования с помощью логистической регрессии для случая непрерывной целевой переменной, принимающей вероятностные значения, а также обобщение формулы веса категории переменной и подсчета индекса Джини.

Цель исследования: 1) определение обобщенных формул для обобщенного логарифма функции правдоподобия, а также его первой и второй производных (вектора градиента и матрицы Гессе) с использованием логистической функции в целях реализации метода Ньютона для численного подсчета параметров регрессии для случая непрерывной целевой переменной, принимающей вероятностные значения; 2) определение обобщенной формулы подсчета веса категории переменной (Weight of Evidence, *WoE*) для случая непрерывной целевой переменной, принимающей вероятностные значения; 3) определение формулы подсчета Джини для случая непрерывной целевой переменной, принимающей вероятностные значения.

ОБОБЩЕНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ, ПОДСЧЕТА ВЕСА КАТЕГОРИИ ПЕРЕМЕННОЙ И ИНДЕКСА ДЖИНИ

Классическая статистическая модель логистической регрессии предполагает использование логит-преобразования [4] от полинома первого порядка — линейной комбинации входящих переменных со свободным членом, который можно представить как скалярное произведение (inner product) вектора коэффициентов логистической регрессии и вектора входящих параметров, дополненного единичной константой в качестве первой координаты (первой входящей переменной). Логит-преобразование скалярного произведения является функцией кумулятивного распределения от скалярного произведения для логистического распределения с нулевым математическим ожиданием и среднеквадратическим отклонением, равным $\pi / \sqrt{3}$ [4]:

$$P(\mathbf{c}, \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{c}, \mathbf{x})}}. \qquad (1)$$

Функция правдоподобия в общем виде для задачи бинарной классификации имеет вид [4]

$$L(\mathbf{c}) = L(\mathbf{c}, X, \mathbf{y}) = \prod_{i:\, y_i = 1} P(\mathbf{c}, \mathbf{x}_i) \prod_{i:\, y_i = 0} (1 - P(\mathbf{c}, \mathbf{x}_i)). \qquad (2)$$

Логарифм этой функции имеет вид [4]

$$\ln L(\mathbf{c}) = \ln L(\mathbf{c}, X, \mathbf{y}) = \sum_{i:\, y_i=1} \ln P(\mathbf{c}, \mathbf{x}_i) + \sum_{i:\, y_i=0} \ln (1 - P(\mathbf{c}, \mathbf{x}_i)). \quad (3)$$

Во многих публикациях для данного логарифма функции правдоподобия также предлагается форма записи в виде одной суммы [4]

$$\ln L(\mathbf{c}) = \ln L(\mathbf{c}, X, \mathbf{y}) = \sum_{i=1}^{n} (y_i \ln P(\mathbf{c}, \mathbf{x}_i) + (1 - y_i) \ln (1 - P(\mathbf{c}, \mathbf{x}_i))). \quad (4)$$

Данная запись предполагает использование двух индикаторных функций — для нуля и единицы:

$$\ln L(\mathbf{c}) = \ln L(\mathbf{c}, X, \mathbf{y}) = \sum_{i=1}^{n} (\mathrm{I}_{\{1\}}(y_i) \ln P(\mathbf{c}, \mathbf{x}_i) + \mathrm{I}_{\{0\}}(y_i) \ln (1 - P(\mathbf{c}, \mathbf{x}_i))). \quad (5)$$

Обобщение логистической регрессии для случая непрерывной целевой переменной y_i, принимающей вероятностные значения, состоит в использовании весовой функции, зависимой от непрерывной целевой переменной:

$$\ln L(\mathbf{c}) = \sum_{i=1}^{n} (w(y_i) \ln P(\mathbf{c}, \mathbf{x}_i) + (1 - w(y_i)) \ln (1 - P(\mathbf{c}, \mathbf{x}_i))). \quad (6)$$

Формула (6) представляет сумму выпуклых комбинаций.

Предлагаемые ограничения на весовую функцию:

$$\begin{cases} w(1) = 1; \\ w(0) = 0; \\ \forall y \in [0;1]:\ w(y) \in [0;1]; \\ \forall y_i \in [0;1],\ \forall y_j \in [0;1]: y_i > y_j \Rightarrow w(y_i) > w(y_j). \end{cases} \quad (7)$$

Первые два ограничения обобщают случай бинарной дискретности. Третье ограничение описывает область определения и область значений весовой функции. Четвертое ограничение обеспечивает монотонность непрерывной весовой функции.

Примером порождаемого весовой функцией семейства логарифмов функции правдоподобия является семейство, порождаемое степенной весовой функцией:

$$w(y) = y^{\alpha}. \quad (8)$$

Достаточным условием выполнения ограничений (7) является область значений параметра

$$\alpha > 0. \quad (9)$$

При $\alpha = 1$ получаем классическую формулу логарифма функции правдоподобия (4):

$$\begin{cases} w(y_i) = y_i; \\ \ln L(\mathbf{c}) = \ln L(\mathbf{c}, X, \mathbf{y}) = \sum_{i=1}^{n} (w(y_i) \ln P(\mathbf{c}, \mathbf{x}_i) + (1 - w(y_i)) \ln (1 - P(\mathbf{c}, \mathbf{x}_i))). \end{cases} \quad (10)$$

Выведем обобщенные аналитические формулы первой и второй производных (вектора градиента и матрицы Гессе) обобщенного логарифма функции правдоподобия (6) с весовой функцией и докажем их применимость для случая непрерывной целевой переменной.

Теорема 1. В логистической регрессии при замене в классической формуле для логарифма функции правдоподобия (4) всех значений целевой переменной y_i на соответствующие значения функции от целевой переменной $w(y_i)$ (в случае обобщения (6) для непрерывной целевой переменной, принимающей вероятностные значения) выполняются следующие условия:

1) матрица Гессе для логарифма функции правдоподобия совпадает с классической матрицей Гессе для случая бинарной логистической регрессии (совпадает с классической матрицей Гессе для логарифма классической функции правдоподобия);

2) в формуле вектора градиента для логарифма функции правдоподобия относительно случая бинарной логистической регрессии (относительно классической формулы вектора градиента для логарифма классической функции правдоподобия) все значения целевой переменной y_i заменяются соответствующими значениями произвольной функции от целевой переменной $w(y_i)$ аналогично исходной замене.

Доказательство. Производная составляющей функции логит-преобразования по вектору $\mathbf{c}$ имеет вид [4]

$$\frac{\partial P(\mathbf{c}, \mathbf{x})}{\partial \mathbf{c}} = \frac{e^{-(\mathbf{c}, \mathbf{x})}}{(1+e^{-(\mathbf{c}, \mathbf{x})})^2}\mathbf{x} = P(\mathbf{c}, \mathbf{x})(1-P(\mathbf{c}, \mathbf{x}))\mathbf{x}\,. \tag{11}$$

Первая производная обобщенного логарифма функции правдоподобия (6) (вектор градиента) имеет вид

$$\frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}\left(w(y_i)\frac{d\ln P(\mathbf{c}, \mathbf{x}_i)}{d\mathbf{c}} + (1-w(y_i))\frac{d\ln(1-P(\mathbf{c}, \mathbf{x}_i))}{d\mathbf{c}}\right), \tag{12}$$

$$\frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}\left(\frac{w(y_i)}{P(\mathbf{c}, \mathbf{x}_i)}\frac{dP(\mathbf{c}, \mathbf{x}_i)}{d\mathbf{c}} + \frac{(1-w(y_i))}{1-P(\mathbf{c}, \mathbf{x}_i)}\frac{d(1-P(\mathbf{c}, \mathbf{x}_i))}{d\mathbf{c}}\right), \tag{13}$$

$$\frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}\left(\frac{w(y_i)}{P(\mathbf{c}, \mathbf{x}_i)}\frac{dP(\mathbf{c}, \mathbf{x}_i)}{d\mathbf{c}} + \frac{(w(y_i)-1)}{1-P(\mathbf{c}, \mathbf{x}_i)}\frac{dP(\mathbf{c}, \mathbf{x}_i)}{d\mathbf{c}}\right), \tag{14}$$

$$\frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}(w(y_i)(1-P(\mathbf{c}, \mathbf{x}_i)) + (w(y_i)-1)P(\mathbf{c}, \mathbf{x}_i))\,\mathbf{x}_i, \tag{15}$$

$$\mathbf{g}(\mathbf{c}) = \frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}(w(y_i)-P(\mathbf{c}, \mathbf{x}_i))\,\mathbf{x}_i\,. \tag{16}$$

Вторая производная обобщенного логарифма функции правдоподобия (6) (матрица Гессе) имеет вид

$$\frac{d^2\ln L(\mathbf{c})}{d\mathbf{c}^2} = \frac{d\sum_{i=1}^{n}(w(y_i)-P(\mathbf{c}, \mathbf{x}_i))\,\mathbf{x}_i}{d\mathbf{c}}, \tag{17}$$

$$H(\mathbf{c}) = \frac{d^2\ln L(\mathbf{c})}{d\mathbf{c}^2} = -\sum_{i=1}^{n}P(\mathbf{c}, \mathbf{x}_i)(1-P(\mathbf{c}, \mathbf{x}_i))\,\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\,. \tag{18}$$

Вторая производная (18) не зависит от целевой переменной и имеет вид классической матрицы Гессе для классической бинарной логистической регрессии [4]. Матрица Гессе является симметричной, как линейная комбинация произведений Кронекера векторов измерений входящих параметров, изначально дополненных единицей.

Доказательство теоремы 1 завершено.

Следствие из теоремы 1. Первая производная (16) при весовой функции, равной непрерывной целевой переменной (10) (например, при единичном значении α семейства степенных функций), соответствует классическому вектору градиента для классической бинарной логистической регрессии [4]:

$$\mathbf{g}(\mathbf{c}\,|\,w(y)=y) = \frac{d\ln L(\mathbf{c})}{d\mathbf{c}} = \sum_{i=1}^{n}(y_i - P(\mathbf{c}, \mathbf{x}_i))\,\mathbf{x}_i\,. \tag{19}$$

Одним из основных выводов является то, что запись классической бинарной регрессии в виде (4) полностью применима для случая непрерывной целевой переменной, принимающей вероятностные значения, и классические формулы (19) и (18) для первой и второй производных (вектора градиента и матрицы Гессе) верны и не требуют изменения при использовании метода Ньютона. При доказательстве формул производных не использовалось ограничение бинарной дискретности.

Вторым важным заключением является возможность обобщения и изменения логарифма функции правдоподобия согласно формуле (6) с помощью весовых функций, что обеспечивает выпуклую комбинацию. Эти функции не влияют на случай бинарной целевой переменной благодаря ограничениям (7). Обобщенная матрица Гессе (18) при использовании весовой функции не изменяется относительно классической, а формула вектора градиента (16) изменяется незначительно относительно классической формулы (19).

Согласно методу Ньютона [4]

$$\mathbf{c}_{k+1} = \mathbf{c}_k - H^{-1}(\mathbf{c}_k)\mathbf{g}(\mathbf{c}_k). \tag{20}$$

Для завершения метода Ньютона [4] требуется выполнение условия

$$||\mathbf{c}_{k+1} - \mathbf{c}_k|| < \varepsilon. \tag{21}$$

Третий важный вывод состоит в возможности приближения коэффициентов предлагаемого непрерывного обобщения логистической регрессии для вероятностной целевой переменной с помощью коэффициентов классической бинарной логистической регрессии при создании искусственной выборки с бинарной целевой переменной в случае, если весовая функция равна непрерывной целевой переменной, принимающей вероятностные значения. При этом используются множественные включения одних и тех же наблюдений с оригинальной выборки в искусственную выборку, но с разными бинарными исходами, пропорционально непрерывным вероятностям бинарных исходов в оригинальной выборке. Этот вывод сформулируем в виде следующей теоремы.

Теорема 2. Коэффициенты бинарной логистической регрессии на искусственной выборке размера mn стремятся к коэффициентам непрерывного обобщения логистической регрессии на оригинальной выборке размера n при $m \to \infty$, если искусственная выборка формируется на основании оригинальной выборки следующим образом:

1) искусственная выборка содержит $m_{0i}(m, y_i)$ вхождений векторов наблюдения $\mathbf{x}_i$ с целевой переменной, равной нулю, и $m_{1i}(m, y_i)$ вхождений с целевой переменной, равной единице; при этом имеем равенство

$$m_{0i}(m, y_i) + m_{1i}(m, y_i) = m \quad \forall i \in \{1, \ldots, n\}; \tag{22}$$

2) выполняется следующее множество условий относительно непрерывной целевой переменной y_i, принимающей вероятностные значения

$$\frac{m_{1i}(m, y_i)}{m} \underset{m\to\infty}{\to} y_i \quad \forall i \in \{1, \ldots, n\}. \tag{23}$$

Два условия выполняются (в обратном порядке), например, при использовании функции целой части:

$$\begin{cases} m_{1i}(m, y_i) = [my_i]; \\ m_{0i}(m, y_i) = m - m_{1i}(m, y_i). \end{cases} \tag{24}$$

Доказательство. Для искусственной выборки классическая функция правдоподобия бинарной логистической регрессии в терминах оригинальной выбор-

ки, используя формулу (2), но с учетом множественных вхождений каждого наблюдения оригинальной выборки в подмножества двух разных классов искусственной выборки, имеет следующий вид:

$$L_{mn}(\mathbf{c}) = \prod_{i\in\{1,\ldots,n\}} (P(\mathbf{c}, \mathbf{x}_i))^{m_{1i}(m,\, y_i)} \prod_{i\in\{1,\ldots,n\}} (1-P(\mathbf{c}, \mathbf{x}_i))^{m_{0i}(m,\, y_i)}, \tag{25}$$

$$L_{mn}(\mathbf{c}) = \prod_{i\in\{1,\ldots,n\}} (P(\mathbf{c}, \mathbf{x}_i))^{m_{1i}(m,\, y_i)} (1-P(\mathbf{c}, \mathbf{x}_i))^{m_{0i}(m,\, y_i)}. \tag{26}$$

Формула корня степени m функции правдоподобия $L_{mn}(\mathbf{c})$ с учетом равенства (22) имеет вид

$$\sqrt[m]{L_{mn}(\mathbf{c})} = \prod_{i\in\{1,\ldots,n\}} (P(\mathbf{c}, \mathbf{x}_i))^{\frac{m_{1i}(m,\, y_i)}{m}} (1-P(\mathbf{c}, \mathbf{x}_i))^{\left(1-\frac{m_{1i}(m,\, y_i)}{m}\right)}. \tag{27}$$

Предел корня степени m функции правдоподобия $L_{mn}(\mathbf{c})$ при $m\to\infty$ с учетом равенства (23) имеет вид

$$\forall \mathbf{c}\in R^C:\ \lim_{m\to\infty} \sqrt[m]{L_{mn}(\mathbf{c})} = \prod_{i\in\{1,\ldots,n\}} (P(\mathbf{c}, \mathbf{x}_i))^{y_i} (1-P(\mathbf{c}, \mathbf{x}_i))^{1-y_i}, \tag{28}$$

$$\forall \mathbf{c}\in R^C:\ \lim_{m\to\infty} \sqrt[m]{L_{mn}(\mathbf{c})} = L(\mathbf{c}), \tag{29}$$

где $L(\mathbf{c})$ — классическая функция правдоподобия. Это обобщенно применяется для случая непрерывной целевой переменной, принимающей вероятностные значения, т.е.

$$L(\mathbf{c}) = L(\mathbf{c}, X, \mathbf{y}) = \sum_{i=1}^{n} (P(\mathbf{c}, \mathbf{x}_i))^{y_i} (1-P(\mathbf{c}, \mathbf{x}_i))^{1-y_i}, \tag{30}$$

что в терминах выражения (10) записывается как

$$\begin{cases} w(y_i) = y_i; \\ L(\mathbf{c}) = L(\mathbf{c}, X, \mathbf{y}) = \sum\limits_{i=1}^{n} (P(\mathbf{c}, \mathbf{x}_i))^{w(y_i)} (1-P(\mathbf{c}, \mathbf{x}_i))^{1-w(y_i)}. \end{cases} \tag{31}$$

Выражение (29) можно переписать следующим образом:

$$\forall \mathbf{c}\in R^C:\quad \ln\left(\lim_{m\to\infty} \sqrt[m]{L_{mn}(\mathbf{c})}\right) = \ln L(\mathbf{c}), \tag{32}$$

$$\forall \mathbf{c}\in R^C:\quad \lim_{m\to\infty} \frac{\ln L_{mn}(\mathbf{c})}{m} = \ln L(\mathbf{c}), \tag{33}$$

т.е. $\forall \mathbf{c}\in R^C$ при достаточно большом значении m

$$\ln L_{mn}(\mathbf{c}) \approx m \ln L(\mathbf{c}), \tag{34}$$

или, более точно, для выражения (33) точечная сходимость формулируется

$$\forall \mathbf{c}\in R^C,\ \forall \varepsilon>0\ \exists m(\mathbf{c},\varepsilon)\in N,\ \forall m > m(\mathbf{c},\varepsilon):\ \left|\frac{\ln L_{mn}(\mathbf{c})}{m} - \ln L(\mathbf{c})\right| < \varepsilon, \tag{35}$$

$$\forall \mathbf{c}\in R^C:\ \left\{\frac{\ln L_{mn}(\mathbf{c})}{m}\right\}_{m=1}^{\infty} \to \ln L(\mathbf{c}), \tag{36}$$

где N — множество натуральных чисел.

Очевидно, что

$$\mathbf{c}^*_{mn} = \operatorname*{agrmax}_{\mathbf{c}\in R^C} L_{mn}(\mathbf{c}) = \operatorname*{agrmax}_{\mathbf{c}\in R^C} \sqrt[m]{L_{mn}(\mathbf{c})} = \operatorname*{agrmax}_{\mathbf{c}\in R^C} \ln \sqrt[m]{L_{mn}(\mathbf{c})}, \tag{37}$$

$$\mathbf{c}^*_{mn} = \operatorname*{agrmax}_{\mathbf{c}\in R^C} \ln L_{mn}(\mathbf{c}) = \operatorname*{agrmax}_{\mathbf{c}\in R^C} \frac{\ln L_{mn}(\mathbf{c})}{m}. \tag{38}$$

Формулы (37) и (38) описывают вектор оптимальных коэффициентов бинарной логистической регрессии. Далее введем обозначение

$$\mathbf{c}_n^* = \underset{\mathbf{c}\in R^C}{\operatorname{agrmax}} L(\mathbf{c}) = \underset{\mathbf{c}\in R^C}{\operatorname{agrmax}} \ln L(\mathbf{c}), \tag{39}$$

а также обозначение подмножества в пространстве R^C:

$$R_r^C = \{\mathbf{c}: \|\mathbf{c}\| \le r\}, \tag{40}$$

$$R_r^C \subset R^C. \tag{41}$$

Тогда выражения (38) и (39) примут вид

$$\mathbf{c}_{mn}^* = \lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \ln L_{mn}(\mathbf{c}) = \lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \frac{\ln L_{mn}(\mathbf{c})}{m}, \tag{42}$$

$$\mathbf{c}_n^* = \lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \ln L(\mathbf{c}). \tag{43}$$

С учетом утверждения (35) введем обозначение

$$m_r(\mathbf{c},\varepsilon) = \max_{R_r^C} m(\mathbf{c},\varepsilon). \tag{44}$$

Аналогично выражению (35) записывается равномерная сходимость на R_r^C:

$$\forall\varepsilon>0 \;\; \exists m_r(\mathbf{c},\varepsilon)\in N, \;\; \forall\mathbf{c}\in R_r^C, \;\; \forall m > m_r(\mathbf{c},\varepsilon): \left|\frac{\ln L_{mn}(\mathbf{c})}{m} - \ln L(\mathbf{c})\right| < \varepsilon, \tag{45}$$

$$\lim_{m\to\infty} \sup_{\mathbf{c}\in R_r^C} \left|\frac{\ln L_{mn}(\mathbf{c})}{m} - \ln L(\mathbf{c})\right| = 0, \tag{46}$$

$$\frac{\ln L_{mn}(\mathbf{c})}{m} \xrightarrow[R_r^C,\, m\to\infty]{} \ln L(\mathbf{c}). \tag{47}$$

Далее, используя (42) и равномерную сходимость на R_r^C при $m\to\infty$ (47)

$$\lim_{m\to\infty} \mathbf{c}_{mn}^* = \lim_{m\to\infty}\lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \frac{\ln L_{mn}(\mathbf{c})}{m} = \lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \lim_{R_r^C,\, m\to\infty} \frac{\ln L_{mn}(\mathbf{c})}{m}, \tag{48}$$

а также правую часть выражения (47) и равенство (43), получаем

$$\lim_{m\to\infty} \mathbf{c}_{mn}^* = \lim_{r\to\infty} \underset{R_r^C}{\operatorname{agrmax}} \ln L(\mathbf{c}) = \mathbf{c}_n^*. \tag{49}$$

Доказательство теоремы 2 завершено.

Весовая функция, которая отлична от классической, т.е. не равна целевой переменной, вносит смещение в прогнозную вероятность при вероятностных значениях целевой переменной, но не вносит никаких изменений для классического бинарного случая благодаря предложенным ограничениям (7). Этот факт можно учитывать при анализе отклоненных заявок (reject inference) [2, 6].

Классическая формула веса категории переменной — категориальной (дискретной) характеристики имеет вид [2]

$$WoE_i = \ln(g_i / b_i). \tag{50}$$

Категориальный показатель g_i — это отношение количества элементов с единичным целевым результатом в сегменте категории с номером i к общему количеству элементов с единичным целевым результатом всех категорий:

$$g_i = \frac{G_i}{\sum_{i=1}^{c} G_i} = \frac{G_i}{G}. \tag{51}$$

Таким образом, оперируем распределением элементов с единичным целевым результатом по дискретным или дискретизированным значениям переменной (категориям переменной), поэтому имеет место равенство

$$\sum_{i=1}^{c} g_i = 1. \tag{52}$$

Аналогично категориальный показатель b_i — это отношение количества элементов с нулевым целевым результатом в сегменте категории с номером i к общему количеству элементов с нулевым целевым результатом всех категорий:

$$b_i = \frac{B_i}{\sum_{i=1}^{c} B_i} = \frac{B_i}{B}. \tag{53}$$

Также оперируем распределением элементов с нулевым целевым результатом по дискретным или дискретизированным значениям переменной (категориям переменной), отсюда имеем

$$\sum_{i=1}^{c} b_i = 1. \tag{54}$$

На основании весов категории переменной и двух распределений, g_i и b_i, подсчитывается показатель значения информации (Information Value, *IV*) — производный от расстояния Кульбака–Лейблера [2, 3]:

$$IV = \sum_{i=1}^{c} (g_i - b_i) \ln\left(\frac{g_i}{b_i}\right) = \sum_{i=1}^{c} (g_i - b_i) WoE_i. \tag{55}$$

Усовершенствование веса категории переменной *WoE* определяется формулой

$$WoE_i = \ln\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}}\right) - \ln\left(\frac{\sum_{j=1}^{n_i} (1 - y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} (1 - y_{ij})}\right). \tag{56}$$

Здесь введена двойная нумерация вероятностной целевой переменной y_{ij}, где индекс i означает номер кластера (всего имеем c кластеров), а индекс j — внутреннюю нумерацию в кластере. Особенность обобщения состоит в использовании сумм вероятностей определенного исхода в кластере, соотнесенной к общей сумме вероятностей определенного исхода. Обобщенное значение информации *IV* также можно записать с использованием данного подхода:

$$IV = \sum_{i=1}^{c} \left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}} - \frac{\sum_{j=1}^{n_i} (1 - y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} (1 - y_{ij})}\right)\left(\ln\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}}\right) - \ln\left(\frac{\sum_{j=1}^{n_i} (1 - y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} (1 - y_{ij})}\right)\right), \tag{57}$$

$$IV = \sum_{i=1}^{c} \left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}} - \frac{\sum_{j=1}^{n_i} (1 - y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} (1 - y_{ij})}\right) WoE_i. \tag{58}$$

Теорема 3. На искусственной выборке размера mn с бинарным исходом, построенной на оригинальной выборке размера n с соблюдением двух условий теоремы 2, при $m \to \infty$ классические веса категорий переменных и значения информации стремятся к предложенным обобщениям (56) и (58) для случая непрерывной целевой переменной, принимающей вероятностные значения.

Доказательство. Форма записи второго условия теоремы 2 при введении двойной индексации (номер сегмента и номер внутри сегмента) имеет следующий вид:

$$\forall i \in \{1, \ldots, c\}, \ \forall j \in \{1, \ldots, n_i\}: \ \frac{m_{1ij}(m, y_{ij})}{m} \underset{m\to\infty}{\to} y_{ij}, \tag{59}$$

а форма записи первого условия теоремы 2 относительно $m_{0ij}(m, y_{ij})$ имеет вид

$$m_{0ij}(m, y_{ij}) = m - m_{1ij}(m, y_{ij}). \tag{60}$$

Классические формулы (51) и (53) для анализа выборки с бинарным исходом в случае искусственной выборки имеют вид

$$g_i(m) = \frac{G_i(m)}{\sum_{i=1}^{c} G_i(m)} = \frac{\sum_{j=1}^{n_i} m_{1ij}(m, y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} m_{1ij}(m, y_{ij})}, \tag{61}$$

$$b_i(m) = \frac{B_i(m)}{\sum_{i=1}^{c} B_i(m)} = \frac{\sum_{j=1}^{n_i} m_{0ij}(m, y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} m_{0ij}(m, y_{ij})}. \tag{62}$$

Разделив числитель и знаменатель выражений (61) и (62) на m с учетом (59) и (60), вычислим пределы выражений (61) и (62):

$$\lim_{m\to\infty} g_i(m) = \frac{\sum_{j=1}^{n_i} \lim_{m\to\infty} \frac{m_{1ij}(m, y_{ij})}{m}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} \lim_{m\to\infty} \frac{m_{1ij}(m, y_{ij})}{m}} = \frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}}, \tag{63}$$

$$\lim_{m\to\infty} b_i(m) = \frac{\sum_{j=1}^{n_i} \left(1 - \lim_{m\to\infty} \frac{m_{1ij}(m, y_{ij})}{m}\right)}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} \left(1 - \lim_{m\to\infty} \frac{m_{1ij}(m, y_{ij})}{m}\right)} = \frac{\sum_{j=1}^{n_i} (1 - y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} (1 - y_{ij})}. \tag{64}$$

Введем обозначения согласно классическим формулам (50) и (55) для искусственной выборки с бинарным исходом

$$WoE_i(m) = \ln\left(\frac{g_i(m)}{b_i(m)}\right), \tag{65}$$

$$IV(m) = \sum_{i=1}^{c} (g_i(m) - b_i(m)) \ln\left(\frac{g_i(m)}{b_i(m)}\right) = \sum_{i=1}^{c} (g_i(m) - b_i(m)) WoE_i(m). \tag{66}$$

Предел веса категории переменной с использованием пределов (63) и (64):

$$\lim_{m\to\infty} WoE_i(m) = \ln\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}}\right) - \ln\left(\frac{\sum_{j=1}^{n_i}(1-y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(1-y_{ij})}\right). \quad (67)$$

Предел значения информации с использованием пределов (63), (64) и (67):

$$\lim_{m\to\infty} IV(m) =$$

$$= \sum_{i=1}^{c}\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}} - \frac{\sum_{j=1}^{n_i}(1-y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(1-y_{ij})}\right)\left(\ln\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}}\right) - \ln\left(\frac{\sum_{j=1}^{n_i}(1-y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(1-y_{ij})}\right)\right), \quad (68)$$

$$\lim_{m\to\infty} IV(m) = \sum_{i=1}^{c}\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij}} - \frac{\sum_{j=1}^{n_i}(1-y_{ij})}{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(1-y_{ij})}\right)\lim_{m\to\infty} WoE_i(m). \quad (69)$$

Таким образом, согласно обозначениям (56) и (57) выполняются равенства

$$\lim_{m\to\infty} WoE_i(m) = WoE_i, \quad (70)$$

$$\lim_{m\to\infty} IV(m) = IV, \quad (71)$$

т.е. классические показатели для искусственной выборки с бинарным исходом сходятся к предлагаемым обобщенным показателям весов категорий переменных и значениям информации на оригинальной выборке с вероятностным исходом. Доказано также выполнимость равенства (58) согласно полученному равенству (69).

Доказательство теоремы 3 завершено.

Основными способами подсчета показателя Джини, используемыми для тестирования и оценки качества модели, являются кривая Лоренца, а также кривая операционной характеристики приемника (Receiver Operating Characteristic curve) [2, 3, 7], которую можно преобразовать в график зависимости значений кумулятивной функции распределения нулевых («плохих») элементов от значений кумулятивной функции распределения единичных («хороших») элементов. Обе кумулятивные функции представлены в виде параметрического графика с параметром уровня порога отсечения для прогнозируемой вероятности модели. Входящим набором является набор двухмерных векторов фактических и прогнозируемых значений (y_i, s_i). Тогда формула, которая обобщается для подсчета индекса Джини, имеет вид [7]

$$GINI = \left(\int_{s\in S} F_B(s)dF_G(s) - \frac{1}{2}\right) \Big/ \left(\frac{1}{2}\right). \quad (72)$$

Далее интеграл можно оценить численно [7]:

$$\int_{s\in S} F_B(s)dF_G(s) = \sum_{s_i\in S} \frac{(F_B(s_i) + F_B(s_{i-1}))}{2}(F_G(s_i) - F_G(s_{i-1})). \quad (73)$$

Обобщение кумулятивных распределений:

$$F_G(t) = \frac{\sum\limits_{i:\, s_i \le t} y_i}{\sum\limits_{i=1}^{n} y_i}, \tag{74}$$

$$F_B(t) = \frac{\sum\limits_{i:\, s_i \le t} (1 - y_i)}{\sum\limits_{i=1}^{n} (1 - y_i)}. \tag{75}$$

Теорема 4. На искусственной выборке размера mn с бинарным исходом, построенной на оригинальной выборке размера n с соблюдением двух условий теоремы 2, при $m \to \infty$ классический индекс Джини дискретной модели бинарного выбора стремится к предложенному обобщению (72)–(75) для модели с непрерывной входящей целевой переменной, принимающей вероятностные значения.

Доказательство. В теоремах 2 и 3 доказана сходимость весов категорий переменных и вектора коэффициентов логистической регрессии, что обусловливает сходимость прогнозируемых значений $s_i(m) \underset{m\to\infty}{\to} s_i$. Поэтому пределы классических эмпирических функций распределения для искусственной выборки с учетом двух условий теоремы 2 имеют следующий вид (условие $\lim\limits_{m\to\infty} s_i(m) \le t$ заменяется множителем — индикаторной функцией неравенства, что подтверждает корректность рассуждений, изложенных ниже):

$$\lim_{m\to\infty} F_G(t, m) =$$

$$= \lim_{m\to\infty} \frac{\sum\limits_{i:\, s_i(m) \le t} m_{1i}(m,\, y_i)}{\sum\limits_{i=1}^{n} m_{1i}(m,\, y_i)} = \frac{\sum\limits_{i:\, \lim\limits_{m\to\infty} s_i(m) \le t} \lim\limits_{m\to\infty} \frac{m_{1i}(m,\, y_i)}{m}}{\sum\limits_{i=1}^{n} \lim\limits_{m\to\infty} \frac{m_{1i}(m,\, y_i)}{m}} = \frac{\sum\limits_{i:\, s_i \le t} y_i}{\sum\limits_{i=1}^{n} y_i},$$

$$\lim_{m\to\infty} F_B(t, m) =$$

$$= \lim_{m\to\infty} \frac{\sum\limits_{i:\, s_i(m) \le t} m_{0i}(m,\, y_i)}{\sum\limits_{i=1}^{n} m_{0i}(m,\, y_i)} = \frac{\sum\limits_{i:\, \lim\limits_{m\to\infty} s_i(m) \le t} \left(1 - \lim\limits_{m\to\infty} \frac{m_{1i}(m,\, y_i)}{m}\right)}{\sum\limits_{i=1}^{n} \left(1 - \lim\limits_{m\to\infty} \frac{m_{1i}(m,\, y_i)}{m}\right)} = \frac{\sum\limits_{i:\, s_i \le t} (1 - y_i)}{\sum\limits_{i=1}^{n} (1 - y_i)}.$$

В терминах обозначений (74) и (75) полученный результат принимает вид

$$\lim_{m\to\infty} F_G(t, m) = F_G(t),$$

$$\lim_{m\to\infty} F_B(t, m) = F_B(t).$$

Как следствие, с учетом всех упомянутых фактов сходится показатель площади под кривой (Area Under Curve, AUC) аналогично интегралу Лебега:

$$AUC(m) = AUC_m = \int\limits_{s(m) \in S(m)} F_B(s(m), m)\, dF_G(s(m), m),$$

$$AUC_m =$$

$$= \sum_{s_i(m)\in S(m)} \frac{(F_B(s_i(m),m)+F_B(s_{i-1}(m),m))}{2}(F_G(s_i(m),m)-F_G(s_{i-1}(m),m)),$$

$$\lim_{m\to\infty} AUC_m = \sum_{s_i\in S} \frac{(F_B(s_i)+F_B(s_{i-1}))}{2}(F_G(s_i)-F_G(s_{i-1})).$$

Тогда согласно (72) сходится последовательность

$$\lim_{m\to\infty} GINI(m) = 2 \lim_{m\to\infty} AUC(m) - 1 = GINI.$$

Доказательство теоремы 4 завершено.

Как указано в доказательстве теоремы 4, в теореме 2 доказывается сходимость относительно целевой переменной в условиях независимости от параметра m для оригинальной матрицы наблюдений, но в условиях использования весов категорий переменных в качестве значений входящих переменных для матрицы наблюдений важен факт сходимости для *WoE*, что доказано в теореме 3. Таким образом, гарантируется сходимость комплексного классического подхода на двух уровнях, когда в качестве входящих переменных в логистической регрессии используется *WoE*. Данный подход включает:

• расчет матрицы наблюдений с помощью *WoE*-преобразований для входящих категориальных (либо дискретизированных на интервалы) переменных с использованием предложенной формулы для вероятностной целевой переменной;

• выполнение моделирования с помощью логистической регрессии с использованием формулы для вероятностной целевой переменной и некоторой весовой функции (например, равной целевой переменной).

ПРИМЕР ПОСТРОЕНИЯ СКОРИНГОВОЙ МОДЕЛИ С ИСПОЛЬЗОВАНИЕМ ВЕСОВОЙ ФУНКЦИИ, РАВНОЙ НЕПРЕРЫВНОЙ ВЕРОЯТНОСТНОЙ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

В рассмотренном ниже примере обучаемая модель предназначена для прогнозирования индикатора отсутствия выхода за более чем 60 дней просрочки за период девяти месяцев наблюдения после даты выдачи для потребительского кредитования. В модель включены следующие категориально-интервальные переменные с соответствующими коэффициентами обобщенной логистической регрессии:

1) отрасль и тип собственности организации трудоустройства ($c_1 = 0{,}508406067253879$);

2) пол и возраст клиента ($c_2 = 0{,}597316083071572$);

3) образование и текущий стаж работы в организации ($c_3 = 0{,}464335289146336$);

4) образование, семейное положение и количество детей ($c_4 = 0{,}493999578172318$);

5) отрасль, должность и общий стаж работы ($c_5 = 0{,}209020022855618$).

Коэффициент смещения модели логистической регрессии: $c_0 = 2{,}65817577386769$. Результаты подсчета обобщенных весов категорий *WoE* даны в таблице (Information Value = 0.261555978700409). Использование вероятностной целевой переменной отвечает анализу отклоненных заявок [2, 6], выполненного с помощью усовершенствованного метода итеративной классификации для вероятностной целевой переменной [6]. При этом только для отклоненных заявок вероятности всегда близки с точностью до 10^{-6} к соответствующим обратным прогнозам на обучающей выборке. Индекс Джини на бинарной тестовой выборке равен 40,11%, а обобщенный индекс Джини на обучающей выборке (обратные прогнозы) равен 40,18%.

Таблица

Переменные (пол, возраст)	Доля выборки, %	Сумма вероятности		Доля (%) суммы вероятности		Значение *WoE*
		наступления дефолта	ненаступления дефолта	наступления дефолта	ненаступления дефолта	
М, ≤25	9,00	2 610,33171	16 511,66829	18,75	8,31	– 0,81336
М, 26÷29 лет	8,76	1 691,91162	16 938,08838	12,15	8,53	– 0,35424
М, 30÷34 лет	8,79	1 479,30018	17 212,69982	10,62	8,66	– 0,20387
М, 35÷41 лет	9,24	1 284,60993	18 348,39007	9,23	9,24	0,00114
М, 42÷49 лет	6,68	737,17629	13 470,82371	5,29	6,78	0,24751
М, >49 лет	6,39	526,14319	13 049,85681	3,78	6,57	0,55301
Ж, ≤27 лет	10,11	2 046,35376	19 444,64624	14,70	9,79	– 0,40643
Ж, 28÷33 лет	9,02	1 153,93200	18 011,06800	8,29	9,07	0,08987
Ж, 34÷40 лет	9,88	980,09372	20 029,90628	7,04	10,08	0,35939
Ж, 41÷47 лет	7,89	600,96158	16 172,03842	4,32	8,14	0,63456
Ж, 48÷54 лет	7,01	440,11943	14 462,88057	3,16	7,28	0,83435
Ж, >54 лет	7,23	372,86867	14 998,13133	2,68	7,55	1,03651
Общий итог:	100%	13 923,80205	198 650,19795	100%	100%	0
		212 574				

ЗАКЛЮЧЕНИЕ

Предложено обобщение метода моделирования с помощью логистической регрессии путем обобщения логарифма функции правдоподобия на непрерывный отрезок [0; 1] действительной оси для целевой переменной, что позволяет использовать вероятностную целевую переменную в методе максимального правдоподобия (Maximum Likelihood Estimation method). Также определено обобщение формулы для подсчета веса категории переменной и индекса Джини для случая вероятностной целевой переменной. Доказано четыре теоремы: 1) о формулах вектора градиента и матрицы Гессе для весовой функции с использованием вероятностной целевой переменной; 2) о возможности приближения коэффициентов логистической модели с вероятностной целевой переменной с помощью модели на специально построенной искусственной выборке с бинарным исходом, если весовая функция равна вероятностной целевой переменной; 3) о сходимости *WoE*-преобразований для входящих категориальных (либо дискретизированных на интервалы) переменных на искусственно построенных приближающих выборках с бинарным исходом; 4) о сходимости индекса Джини на искусственно построенных приближающих выборках с бинарным исходом. Следствие первой теоремы окончательно подтверждает применимость классического метода Ньютона (включая классические формулы вектора градиента и матрицы Гессе) без изменений при обобщении классической модели логистической регрессии на вероятностную целевую переменную, используя весовую функцию, равную целевой переменной.

Предложенные обобщения имеют существенные преимущества перед классическими формулами подсчета. Главным преимуществом и отличием от классического случая является возможность использования вероятностной целевой переменной либо непрерывной целевой переменной другой природы [5], принимающей значения с интервала 0% ÷100%, например для моделирования показателя относительных потерь, причиняемых дефолтом (Loss Given by Default). Преимуществом также является обобщение всего процесса моделирования для непрерывной целевой переменной — подготовки входящих значений переменных в виде обобщенного преобразования в вес категории переменной, подсчета коэффициентов обобщенной логистической регрессии, оценки качества разработанной модели с помощью обобщенного индекса Джини. Кроме того, имеются доказательства классических формул для метода максимального правдоподобия, веса категории переменной и индекса Джини для бинарной целевой переменной, как частного случая обобщенных формул для целевой переменной, принимающей вероятностные значения.

Одним из основополагающих следствий обобщения формулы веса категории переменной является обобщение показателя значения информации IV. Важным следствием обобщения логарифма функции правдоподобия является возможность использования разнородных непрерывных весовых функций, которые при введенных ограничениях приравнивают логарифм функции правдоподобия к классическому значению на множестве бинарных значений целевой переменной.

Введенные усовершенствования позволяют решать задачи вероятностного моделирования при нечеткой бинарной классификации входящих данных, в частности более эффективно решать задачи включения и анализа отклоненных заявок (reject inference) [5, 6], как частично классифицированных выведенных данных (inferred data) в кредитном скоринге, а также выполнять моделирование показателей, принимающих значения с интервала 0% ÷100% . Классическим примером является задача моделирования относительных потерь, причиняемых реализацией события дефолта (Loss Given by Default).

Основными направлениями перспективных исследований являются более глубокое изучение степеней влияния различных типов предложенной весовой функции в формуле логарифма функции правдоподобия и обобщение других методов категориальной регрессии.

СПИСОК ЛИТЕРАТУРЫ

1. Лобанова А.А., Чугунова А.В. Энциклопедия финансового риск-менеджмента. — М.: Альпина Паблишер, 2003. — 786 с.
2. Siddiqi Naeem. Credit risk scorecards: developing and implementing intelligent credit scoring. — Hoboken: John Wiley & Sons, Inc., 2006. — 196 p.
3. Thomas C. Lyn, Edelman B. David, Crook N. Jonathan. Credit scoring and its applications. — Philadelphia: Society for Industrial and Applied Mathematics, 2002. — 248 p.
4. Allison D. Paul. Logistic regression using the SAS® System: Theory and Application. — Cary: SAS Institute Inc., 1999. — 287 p.
5. Мэйз Э. Руководство по кредитному скорингу. — Минск: Гревцов Паблишер, 2008. — 464 с.
6. Солошенко О.М. Вдосконалення методу ітеративної класифікації з включення відхилених заявок у кредитному скорингу // Наукові вісті НТУУ «КПІ». — 2014. — № 5. — С. 63–69.
7. Солошенко О.М. Спосіб розрахунку показника Джині, статистики Колмогорова–Смирнова та відстані Махаланобіса у кредитному скорингу засобами мови SQL // Наукові вісті НТУУ «КПІ». — 2015. — № 1. — С. 29–35.
8. Терентьев А.Н., Бидюк П.И. Метод вероятностного вывода в байесовских сетях по обучающим данным // Кибернетика и системный анализ. — 2007. — № 3. — С. 93–99.

Поступила 17.09.2014