УДК 519.86 DOI: 10.18698/2309-3684-2019-4-100116

Задачи идентификации индивидуальных покупателей на основе анализа больших объемов панельных данных о кассовых чеках

© Ю.И. Димитриенко, А.В. Котельникова

МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

Сформулированы задачи идентификации индивидуальных покупателей на основе анализа больших объемов данных о кассовых чеках в крупном супермаркете. Разработаны модели поведения различных категорий индивидуальных покупателей в супермаркете. Предложен вычислительный алгоритм решения задач идентификации индивидуальных покупателей по панельным данным кассовых чеков. Алгоритм является универсальным, так как не использует никаких персональных данных о покупателе, а построен на анализе только покупательской активности, вычисляемой на основе панельных данных о кассовых чеках. Алгоритм позволяет идентифицировать группы покупателей, а также с определенной вероятностью, отдельного индивидуального покупателя. В качестве примера применения разработанных моделей и вычислительных алгоритмов использовались товарные чеки из сети супермаркетов компании X5Retail Group за некоторый промежуток времени.

Ключевые слова: *индивидуальный покупатель, задача идентификации, кластеризация, кассовые чеки, анализ больших данных, панельные данные, прогнозирование*

Введение. Для выявления целевой аудитории в крупных супермаркетах в настоящее время активно применяются методы обработки больших массивов данных [1–5]. Как правило, эти методы относятся к математической статистике о продажах конкретных видов товаров и прогнозе динамике изменения продаж [6–20]. Эти методы носят интегральный характер и не позволяют проводить глубокий анализ изменения индивидуальной покупательской активности, и учи-тывать особенности поведения разных покупательских групп в супермаркете, их различную реакцию на проведение акций в супермаркете. Для решения этих задач в работах [21–24] были предложены принципиально новые алгоритмы, основанные на применении методов континуальной механики многомерных сплошных сред. Эти методы позволяют с достаточно хорошей точностью прогнозировать поведение покупателей в течение длительного периода времени.

Однако у этого метода существует ограничение — для него необходима информация о покупках индивидуальных покупателей в супермаркете в течение определенного периода. Такая информация доступна, например, в интернет магазинах или при продажах владельцам именных карт – банковских, кредитных, скидочных и

накопительных. В этом случае имеется возможность анализировать данные о покупках индивидуального покупателя в течение длительного периода времени. Значительно более сложной является ситуация, когда доступна информация только о кассовых чеках, которые являются безымянными. В этом случае возникает очень сложная задача идентификации индивидуального покупателя в различные моменты времени (дни продаж) на основе обработки только информации из кассовых чеков. Математических постановок этой задачи, а также методов ее решения в настоящее время не известно.

Отметим, что задача идентификации отдельного покупателя и группы покупателей на основе данных из кассовых чеков сама по себе представляет интерес, поскольку позволяет выявлять целевые группы для различных категорий товаров. Покупателей, расплачивающихся в супермаркетах наличными и не использующих карты магазинов, достаточно много — примерно половина от общего числа покупателей. Создание алгоритма идентификации индивидуальных покупателей по безымянным кассовым чекам позволит магазину выяснить, например, на какую категорию покупателей нужно акцентировать внимание в зависимости от того, у какой категории индивидуальных покупателей улучшаются или ухудшаются продажи.

Математическая модель покупок индивидуальными покупателями. В качестве индивидуального покупателя в супермаркете будем считать одного человека, который приходит в магазин с некоторой периодичностью и совершает разовую покупку, о которой свидетельствует один кассовый чек. В один конкретный день можно зафиксировать всех покупателей по кассовым чекам, однако в другие дни имеется просто массив кассовых чеков, про которые неизвестно, принадлежат ли они кому-то из покупателей первого из рассматриваемых дней продаж, и кому именно, если в разные дни в данный магазин ходят одни и те же покупатели, или только часть их.

Предположим, что имеются массивы данных из кассовых чеков некоторого магазина за некоторый период времени. Каждый кассовый чек содержит следующую информацию: номер магазина, дата чека, номер чека, код товара, наименование товара, количество товара, цена за товар в одном экземпляре.

Введем следующие понятия и обозначения:

N — общее количество различных наименований товаров в магазине;

i — глобальный номер отдельного товара в единой нумерации $i = 1,..., N$;

c_i — цена i-го товара в одном экземпляре;

K — количество дней, в течение которых осуществляется анализ и прогнозирование покупок (имеются данные кассовых чеков);

t_k — моменты времени (конец каждого рабочего дня), для которых имеются данные кассовых чеков, $k = 1,...,K$;

M_k — число кассовых чеков в момент времени t_k;

j — локальный номер кассового чека для данного дня t_k, $j = 1,...,M_k$;

M — общее количество кассовых чеков за K наблюдаемых дней:

$$M = \sum_{k=1}^{K} M_k \tag{1}$$

$p_{ji}(t_k)$ — количество купленного i-го товара одним покупателем в j-ом кассовом чеке в момент времени t_k.

Объединяя матрицы все $p_{ji}(t_k)$ покупок за t_k-ый день, сформируем матрицы покупок за все время наблюдений $p_{ji}(t_k)$

$$\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & p_{ji} & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MN} \end{pmatrix}, \tag{2}$$

где $i = 1,...,N$, $j = 1,...,M$, $k = 1,...,K$.

Введем укрупненные группы (категории) товаров (УГТ), которые продаются в рассматриваемом супермаркете. Для продовольственных товаров супермаркета, например, можно ввести следующие УГТ: молочная гастрономия, бакалея, мясная гастрономия, хлебные изделия и т. д.

Обозначим также:

L — число введенных УГТ;

m — номер каждой УГТ, $m = 1,...,L$;

N_m — число наименований товаров в m-й УКГ, т.е.

$$i = i_{m-1} + 1,...,i_m,$$

где $i_m = N_1 + N_m$.

Упорядочим товары $i = 1,...,N$ таким образом, что товары, принадлежащие к одной и той же УКГ, следуют друг за другом, не перемешиваясь с товарами других УГТ, тогда матрицу покупок $p_{ji}(t_k)$ можно представить, как блочную матрицу, состоящую из блоков, представленную в табл. 1.

Таблица 1

Структура матрицы покупок $p_{ji}(t_k)$

№ чека j	УГТ 1 p_1			УГТ 2	УГТ m p_m			УГТ L p_L		
1	p_{11}	…	p_{1i_1}		$p_{1,i_{m-1}+1}$	…	p_{1i_m}	$p_{1,i_{L-1}+1}$	…	p_{1i_L}
2										
…										
j	p_{j1}	…	p_{ji_1}		$p_{j,i_{m-1}+1}$	…	p_{ji_m}	$p_{j,i_{L-1}+1}$	…	p_{ji_L}
…										
M_k	p_{M_k1}	…	$p_{M_ki_1}$		$p_{M_k,i_{m-1}+1}$	…	$p_{M_ki_m}$	$p_{M_k,i_{L-1}+1}$	…	$p_{M_ki_L}$

При этом выполняется условие согласования индексов матриц

$$i_L = N . \tag{3}$$

Составим на основе матрицы покупок $p_{ji}(t_k)$ еще одну матрицу — матрица покупок одного покупателя в m-ой УГТ $\hat{p}_{jm}$:

$$\hat{p}_{jm} = \sum_{i=i_{m-1}+1}^{i_m} p_{ji}, \ \ j = 1,...,M_k, \ \ m = 1,...,L, \ \ k = 1,...,K . \tag{4}$$

То есть в матрице $\hat{p}_{jm}$ все покупки в рамках УГТ отождествляются и суммируются. Введём новое обозначение в виде координатного столбца для кассового чека (КЧ) в заданный момент времени t_k

$$\mathbf{p}_j(t_k) = \left(p_{j1}(t_k),..., p_{jN}(t_k)\right), \ \ j = 1,...,M_k, \ \ k = 1,...,K . \tag{5}$$

Соответствующее обозначение для КЧ по отношению к УГТ будет иметь вид:

$$\hat{\mathbf{p}}_j(t_k) = \left(\hat{p}_{j1}(t_k),..., \hat{p}_{jL}(t_k)\right), \ \ j = 1,...,M_k, \ \ k = 1,...,K . \tag{6}$$

Обозначим Q — число различных индивидуальных покупателей, совершающих покупки в исследуемом супермаркете за весь исследуемый период времени $0 < t < t_k$. То есть один и тот же покупатель, совершающий покупки в разные дни, в этом множестве учитывается только один раз. Номер покупателя в этом общем списке обозначим как s: $s = 1...Q$. Момент времени первой покупки s-го покупателя в супермаркете от начала наблюдения обозначим как T_s^0. Очевидно, что $T_s^0 \le t_k$.

Индивидуальных покупателей разделим на 3 группы:

а) посещающие данный супермаркет регулярно с некоторым среднестатистическим периодом T_p, в течение всего исследуемого периода времени t_k;

б) посещающие данный магазин регулярно, но в течение не всего периода $t_1 \le t_k \le t_K$, а только его части;

в) случайные покупатели, посещающие данный магазин разово или от случая к случаю.

Покупатели, которые посещают магазин регулярно, но с перерывами, рассматриваются как разные покупатели из группы б).

Введем далее основное допущение модели: каждый индивидуальный покупатель посещает данный магазин не чаще, чем 1 раз в день.

Тогда на рассматриваемом интервале времени наблюдения $0 \le t \le t_K$ можно каждому s-ому индивидуальному покупателю поставить в соответствие единственным образом кассовый чек $\mathbf{p}_j\left(T_s^0\right)$ его первой покупки в данном супермаркете в некоторый момент времени T_s^0. Поскольку в каждый день имеется определенная упорядоченность КЧ по их идентификационным номерам, то можно ввести единую порядковую нумерацию кассовых чеков ИП $s = 1...Q$, ставя вперед покупателя с более ранней датой T_s^0 их первой покупки и меньшим идентификационным номером. Перенумерованный таким образом набор всех КЧ, соответствующий упорядоченному списку ИП, обозначим как $\mathbf{q}_s$, $s = 1,2,3,...,Q$. Задание списка ИП $\mathbf{q}_s$ реализуем с помощью линейного матричного преобразования

$$\mathbf{q}_s = \sum_{j=1}^{M_k} B_{sj}\left(t_k\right)\mathbf{p}_j\left(t_k\right),\ k = 1,...,S,\ s = 1,2,3,...,Q,\ S \le K. \qquad (7)$$

Элементами матриц соответствия $B_{sj}\left(t_k\right)$ являются 0 или 1

$$B_{sj}\left(t_k\right) \in \{0,1\}. \tag{8}$$

Элемент матрицы $B_{sj}\left(t_k\right)$ равен 1, если s-ому ИП принадлежит первый полученный им КЧ в момент времени $T_s^0 = t_k$ с номером j. Во всех остальных случаях $B_{sj}\left(t_k\right) = 0$.

Формулировки математических задач идентификации ИП. Первая (вспомогательная) задача идентификации заключается в том, чтобы, зная набор КЧ $\mathbf{p}_s\left(t_k\right)$ на промежутке $t_1 \le t_K \le t_S$, $S \le K$, найти:

- число Q ИП;
- моменты времени T_s^0 первой покупки для каждого ИП;
- матрицы соответствия $B_{sj}\left(t_k\right)$ между КЧ и набором ИП.

Решение этой задачи позволяет найти набор ИП, и определить векторы $\mathbf{q}_s$ (переупорядоченные единым списком первые КЧ).

Основная задача идентификации заключается в том, чтобы установить соответствие между КЧ $\mathbf{p}_j\left(t_k\right)$ и ИП $\mathbf{q}_s$ для произвольных моментов времени $t_s < t_k \le t_K$. Это соответствие также будем описывать матричным преобразованием

$$\mathbf{p}_j\left(t_k\right) = \sum_{s=1}^{Q} A_{js}\left(t_k\right)\mathbf{q}_s, \;\; j = 1,\ldots,M_k, \;\; k = S+1,\ldots,K. \tag{9}$$

В этой задаче необходимо найти матрицу соответствия $A_{js}\left(t_k\right)$ между ИП и соответствующими КЧ.

Элемент матрицы $A_{js}\left(t_k\right)$ равен единице, если КЧ $\mathbf{p}_j\left(t_k\right)$ принадлежит данному ИП $\mathbf{q}_s$, и нулю, если не принадлежит, то есть:

$$A_{js}\left(t_k\right) \in \{0,1\}. \tag{10}$$

Алгоритм решения первой задачи идентификации в начальном приближении. В силу допущения, что в течение дня покупатель не может посетить данный магазин более одного раза, можно сделать вывод, что если какой-либо КЧ принадлежит одному индивидуальному покупателю, то больше никакие чеки в данный момент времени не могут соответствовать этому же индивидуальному покупателю.

Выберем все КЧ в начальный момент времени t_1 за часть индивидуальных покупателей. Следовательно, в начальный момент времени t_1 легко находим матрицу соответствия $B_{sj}\left(t_1\right)$, которая будет равняться единичной матрице:

$$\left(B_{sj}\left(t_1\right)\right)=\begin{pmatrix} E_{M_1} \\ 0 \end{pmatrix},\ \ s=1,...,Q,\ \ j=1,...,M_1, \tag{11}$$

где $\left(E_{M_1}\right)=\left(\delta_{sj}\right)$ — единичная матрица размером M_1, а 0 — означает, что все остальные элементы матрицы — нулевые, δ_{sj} — символ Кронекера.

Легко находим также часть массива $T_s^0 : T_s^0 = t_1$ при $s=1,...,M_1$.

Таким образом, установлено соотношение между КЧ для t_1 и частью индивидуальных покупателей $\mathbf{q}_s$, $s=1...M\left(t_1\right)$, что и будет является частичным решением поставленной первой задачи идентификации индивидуальных покупателей по панельным данным кассовых чеков. Это начальное приближение.

Алгоритм решения первой задачи на последующих приближениях. Для того чтобы найти остальных ИП $\mathbf{q}_s$, $s=M_1+1,...,Q$, в рамках первой задачи идентификации рассмотрим несколько последующих моментов времени t_k, $k=1,2,3,...,S$ $\left(S<K\right)$ Построим алгоритм анализа КЦ, который позволяет найти время $T_p=t_S$ совершения покупок, в течение которого магазин посещают все ИП с момента начала отсчета, причем хотя бы один ИП в течение этого времени посещает магазин только один раз.

Разделим множество всех КЧ $\mathbf{p}_j\left(t_k\right)$ в момент времени t_k, $k=2,3,...,S$ на группы (кластеры). Для этого введем суммарную стоимость одного КЧ с номером j:

$$F_j(t_k)=\sum_{i=1}^{N} p_{ji}(t_k)c_i\ . \tag{12}$$

Кластеры КЧ выделяем по суммарной стоимости покупок в одном КЧ с номером j следующим образом. Сначала находим максимальную сумму покупок в момент времени t_k

$$F_{\max}(t_k)=\max_j F_j(t_k)\,. \tag{13}$$

Далее разделим $F_{\max}(t_k)$ на n интервалов и введем приращение стоимости при переходе от интервала к интервалу: $\Delta F=F_{\max}(t_k)/n$. Подсчитываем, количество p_r КЧ $\mathbf{p}_j\left(t_k\right)$, которое попадает в каждый r-ый интервал по величине функции F — стоимости покупки в одном КЧ. В результате получим кусочно–непрерывную функцию

$p(F)$, которую можно рассматривать как гистограмму $p(F)$. Определим разностную производную от этой функции по следующей формуле:

$$\dot{p}(F_{(r)}) = \frac{1}{\Delta F}\left(p\left(F_{(r+1)}\right) - p\left(F_{(r)}\right)\right),\ r = 1,...,n-1, \tag{14}$$

где $F_{(r)} = r\Delta F$. Примем следующее решающее правило определения границ кластеров покупателей (КЧ): считаем, что точка $F_{(r)}$ является границей кластера с номером m, если знаки производных $\dot{p}\left(F_{(r)}\right)$ и $\dot{p}\left(F_{(r-1)}\right)$ слева и справа в этой точке — различны. Обозначим определенные таким образом границы кластеров как $F_{\langle m\rangle}(t_k)$, $m = 1,...,R$, где R — число кластеров. Очевидно, выполняются условия

$$F_{\langle 1\rangle} = F_{(1)},\ F_{\langle R+1\rangle} = F_{(n)},\ 1 \le R \le n-1. \tag{15}$$

Таким образом, для каждого момента времени t_k и каждого разбиения n определяем число $R(t_k, n)$ кластеров покупателей (КП).

Сравним теперь границы кластеров $F_{\langle m\rangle}(t_k)$ между собой в разные моменты времени, при этом интервалы по стоимости покупок ΔF для разных t_k полагаем одинаковыми. Если границы кластеров $F_{\langle m\rangle}(t_k)$ для разных t_k $k = 2,3,...,S$ при фиксированном n оказываются одинаковыми, то считаем, что алгоритм этой задачи завершен. Если число кластеров различно, то изменяем число n, причем для всех моментов времени — одинаково. Цель алгоритма — получить такое разбиение на кластеры, чтобы во все дни они были примерно одинаковыми, или появилось близкое разделение на кластеры множества всех КЧ с периодом T_p в несколько дней. Сравнение кластеров в различные моменты времени осуществляем с помощью сравнения границ кластеров, в результате получаем следующую задачу многокритериальной минимизации функционала отклонения границ кластеров

$$\begin{gathered}\Delta(t_k, n) = \sum_{m=1}^{R(t_1,n)} (F_{<m>}(t_k) - F_{<m>}(t_1))^2, \\ \Delta(t_k, n) \to \min,\ t_k \to \min.\end{gathered} \tag{16}$$

Условие $t_k \to \min$ означает, что ищется минимальное значение момента времени $t_k = T_p$, при котором функционал $\Delta(t_k, n) \to \min$ имеет минимум.

После решения задачи (16), и нахождения времени T_p и числа кластеров $R\left(T_p, n\right)$, проверяется условие периодичности

$$R\left(t_k, n\right) = R\left(t_k + T_p, n\right), \tag{17}$$

Найденные значения T_p, n и $R\left(T_p, n\right)$ завершают решение первой задачи идентификации ИП.

В качестве оставшейся части $\mathbf{q}_s$, $s > M\left(t_1\right)$ принимаются покупатели, КЧ которых соответствуют всем моментам времени t_k начального периода покупок: $t_1 < t_k < T_p$. Матрицы $B_{sj}(t_k)$ в этом случае состоят из единичных матриц E_{M_k} размером M_k

$$\left(B_{sj}\left(t_k\right)\right) = \begin{pmatrix} 0 \\ E_{M_k} \\ 0 \end{pmatrix} \quad s = 1,...,Q, \quad j = 1,...,M_1, \quad k = 2,3,...,S-1, \tag{18}$$

где 0 – означает нулевые блоки элементов матрицы, а $t_{K_1} = T_p$.

Общее число ИП Q вычисляется по формуле

$$Q = M\left(t_1\right) + ... + M\left(t_{S-1}\right). \tag{19}$$

Алгоритм решения основной задачи идентификации. Для решения основной задачи идентификации введём многомерное евклидово пространство укрупненных групп товаров V_L и расширенное пространство $V_{L+1} = V_L \times \left[t_1, t_K\right]$. Каждый j-ый КЧ $\mathbf{p}_j(t_k)$ в момент времени t_k в этом пространстве V_{L+1} изображается точкой $\hat{\mathbf{p}}_j\left(t_k\right)$, связанная с этим ИП соотношением (7) для некоторого t_k, $k \in \left\{1,..., S\right\}$.

Разделим всех ИП в начальные моменты времени t_k, $k \in \left\{1,..., S\right\}$ на 2 группы: а) и б). Если в чеке $\hat{\mathbf{p}}_j(t_k)$ количество купленных товаров (число ненулевых элементов $p_{ji}(t_k)$, $i = 1,..., N$) меньше некоторого числа: $\sum_{i=1}^{N} p_{ji}(t_k) < p_{\min}$, то будем считать, что это случайный

покупатель из группы б). Остальных покупателей отнесем к группе а).

Рассмотрим для каждого покупателя из группы а) промежуток времени T_p — минимальный период (среднее количество дней, через которое предположительно придёт данный покупатель в данный магазин в следующий раз). Осуществим прогноз числа покупок ИП в моменты времени $t = t_k + wT_p$, $k \in \{1,...,S-1\}$, w — число периодов, $w = 1,2,3,....$ Будем считать, что переход ИП $\mathbf{q}_s$ из начальной точки $\hat{\mathbf{p}}_j(t_k)$ в точку $\hat{\mathbf{p}}_j(t)$ в момент времени t происходит под действием случайных факторов, подчиняясь многомерному нормальному распределению. Введем случайное варьирование точки $\hat{\mathbf{p}}_j(t_k)$ т.е. рассмотрим вектор

$$\hat{\mathbf{p}}_j^{v}(t_k) = \hat{\mathbf{p}}_j(t_k) + \tilde{\tilde{\mathbf{p}}}_j, \tag{20}$$

где $\tilde{\tilde{\mathbf{p}}}_j$ — случайный вектор, который имеет нормальную плотность распределения с нулевым математическим ожиданием:

$$f(\tilde{\tilde{\mathbf{p}}}_j) = \frac{1}{(2\pi)^{N/2}\sigma}\exp\left(-\frac{1}{2\sigma}\left|\tilde{\tilde{\mathbf{p}}}_j\right|^2\right), \tag{21}$$

где $\left|\tilde{\tilde{\mathbf{p}}}_j\right|^2$ — квадрат длины вектора, σ — дисперсия распределения случайной величины.

Рассмотрим в момент времени $t > t_k$ такой КЧ $\mathbf{p}_{j'}(t)$, для которого точка $\hat{\mathbf{p}}_{j'}(t)$ находится на минимальном расстоянии от точки $\hat{\mathbf{p}}_j^{v}(t_k)$, т.е. удовлетворяет условию

$$d(\hat{\mathbf{p}}_{j'}(t), \hat{\mathbf{p}}_j^{v}(t_k)) \to \min \tag{22}$$

среди всех КЧ в данный момент времени t, здесь введено расстояние между точками

$$d_{jj'} = d(\hat{\mathbf{p}}_{j'}, \hat{\mathbf{p}}_j^{v}) = \left|\hat{\mathbf{p}}_{j'} - \hat{\mathbf{p}}_j^{v}\right|, \tag{23}$$

где $\left|\hat{\mathbf{p}}_{j'} - \hat{\mathbf{p}}_j^{v}\right|$ — длина вектора.

Такой КЧ $\mathbf{p}_{j'}(t)$ будем полагать решением основной задачи идентификации ИП $\mathbf{q}_s$ в произвольный момент времени $t = t_k + wT_p$.

Сопоставляя значения $\mathbf{q}_s$ и $\mathbf{p}_{j'}(t_k)$, $k = S+1,...,K$, $s = 1,2,3,...,Q$, находим ненулевое значение матрицы $A_{js}(t_k)$ в уравнении (9).

Результаты численного моделирования. В качестве примера применения разработанных алгоритмов решения задач идентификации были использованы КЧ из 3-х магазинов сети супермаркетов «Перекресток» за некоторый промежуток времени — 100 дней ($K = 100$) за 2012 год, предоставленные компанией «X5 RetailGroup». Было выбрано 18 УГТ ($L = 18$): «алкоголь», «бакалея», «бытовая техника», «бытовая химия», «детское питание», «диабетические продукты», «замороженные продукты», «кондитерские изделия», «молочная гастрономия», «мясная гастрономия», «овощи», «рыба и морепродукты», «салатное производство», «соки, воды, пиво», «табак», «фрукты», «хлебные изделия», «прочее».

Для решения первой задачи идентификации была введена функция распределения $p(F)$ покупателей по сумме покупок в начальный момент времени t_1 при числе разбиений $n = 20$.

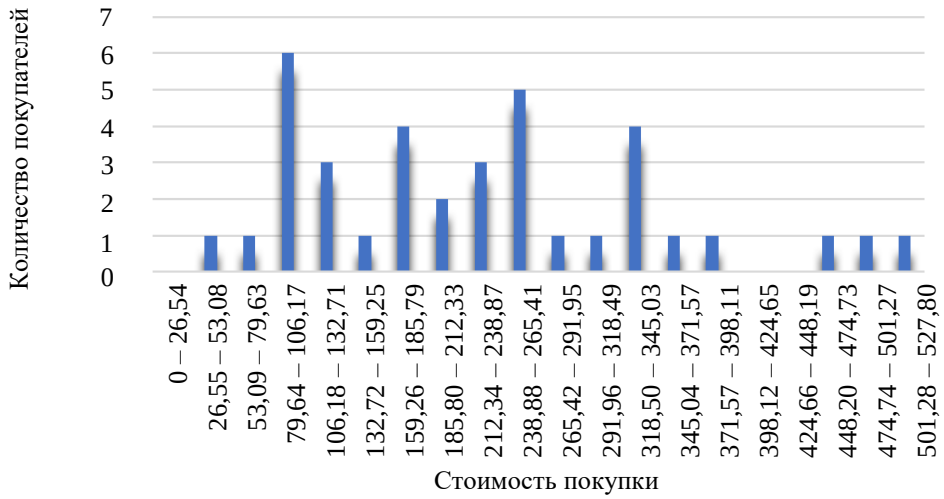

Рис. 1. Гистограмма покупок (КЧ) $p(F)$ в начальный момент времен t_1 для магазина № 1

С помощью формулы (14) был построен график разностной производной $\dot{p}\left(F_{(r)}\right)$ (рис. 2), соответствующий гистограмме $p(F)$, изображенной на рис.1.

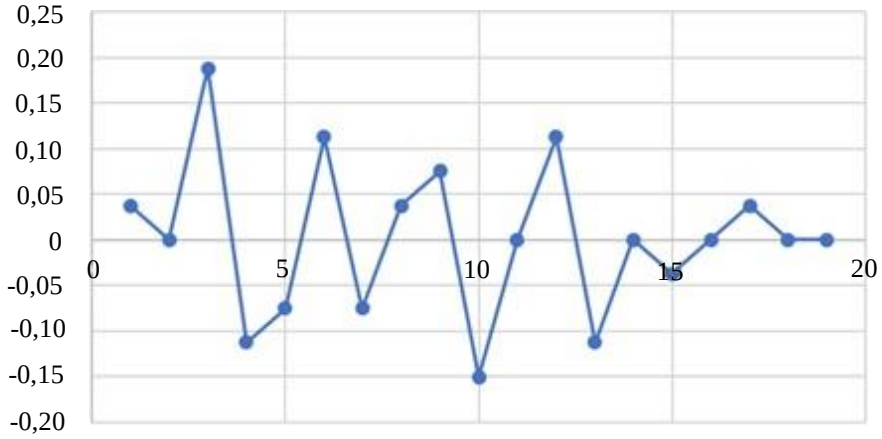

Рис. 2. Разностная производная $\dot{p}\left(F_{(r)}\right)$ функции $p(F)$ в начальный момент времени t_1 для магазина № 1

Далее решаем задачу (16) — поиска числа кластеров $R\left(t_k, n\right)$ и периода T_p. Значение максимальной стоимости КЧ в начальный момент t_1 составляло $F_{\max}(t_1) = 5500$ руб. Кластеры ИП строили по шкале стоимости КЧ. При различных значениях числа разбиений n получалось разное число кластеров $R\left(t_k, n\right)$ для магазина № 1 приведены в табл.2.

Таблица 2

Число кластеров $R\left(t_k, n\right)$ для магазина № 1 при различных значениях n и t_k

t_k	$n=20$	$n=18$	$n=16$	$n=10$	$n=6$
t_1	9	6	10	6	2
t_2	1	1	1	1	1
t_3	1	1	1	3	1
t_4	5	3	2	1	3
t_5	4	4	6	1	1
t_6	6	3	5	1	1
t_7	4	6	3	3	3

Решение задачи оптимизации (16) для магазина № 1 дает следующее значение числа кластеров, которое сохраняется при периодических сдвигах по времени в течение первых 15 дней от момента начала

$$R\left(t_1, 6\right) = R\left(t_1 + T_p, 6\right) = 3, \;\; n = 6.$$

Средний период T_p для магазина № 1 был получен равным T_p=3, причем этот период практически не зависит от числа разбиений n. График зависимости $R(t_k,6)$ приведен на рис. 3. После 15 дней число кластеров $R(t_k,6)$ уменьшается: при $15 \le t_k \le 25$ оно равно 1, а при $t_k \ge 25$ получаем, что $R(t_k,6)=R(t_k+T_p,6)=2$, но средний период T_p для всех рассмотренных моментов времени равен 3.

Заключение. Сформулированы задачи идентификации индивидуальных покупателей на основе анализа больших объемов данных о кассовых чеках в крупном супермаркете. Предложены вычислительные алгоритмы решения задач идентификации индивидуальных покупателей. Алгоритмы позволяют идентифицировать группы покупателей, а также с определенной вероятностью, отдельного индивидуального покупателя. В качестве примера применения разработанных моделей и вычислительных алгоритмов использовались кассовые чеки из сети супермаркетов компании X5Retail Group за некоторый промежуток времени.

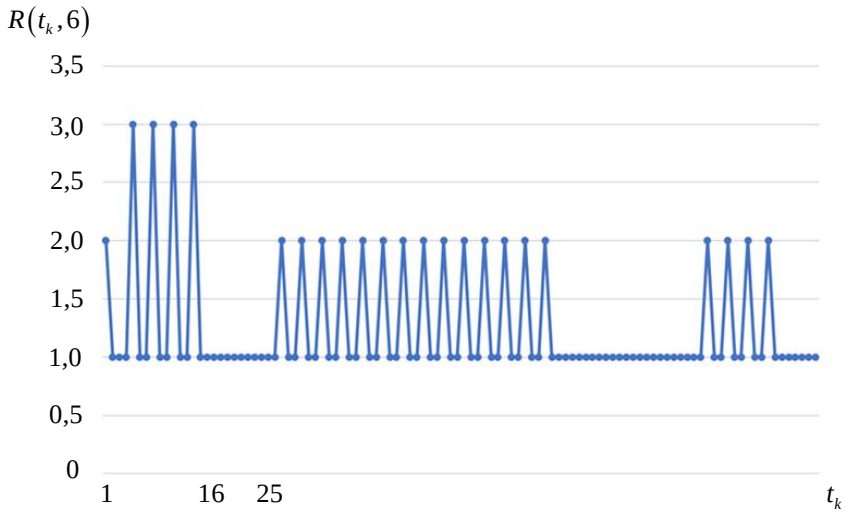

Рис. 3. График изменения количества кластеров $R(t_k,6)$ при $n=6$ для магазина № 1

Показано, что допущения, введенные в модели: относительно существования устойчивых кластеров покупателей, относительно существования среднего периода покупки — являются адекватными.

ЛИТЕРАТУРА

[1] Барсегян А.А., Куприянов М.С., Холод И.И. Тесс М.Д. Елизаров С.И. *Анализ данных и процессов: учебное пособие, 3-е изд*. СПб., БХВ-Петербург, 2009, 512 с.

[2] Демидова Л.А., Кираковский В.В., Пылькин А.Н. *Принятие решений в условиях неопределенности*. Москва, Горячая линия — Телеком, 2012, 288 с.

[3] Демин И.С. *Кластеризация как инструмент интеллектуального анализа данных. Ч. 1: Новые информационные технологии в образовании*. Москва, 1 С-Паблишинг, 2011, с. 98–103.

[4] Журавлев Ю.И., Рязанов В.В., Сенько О.В. *Распознавание. Математические методы. Программная система. Практические применения.* Москва, Изд-во Фазис, 2006, 176 с.

[5] Кулаичев А.П. *Методы и средства комплексного анализа данных: учебное пособие, 4-е изд., перераб. и доп.* Москва, ФОРУМ, Инфра-М, 2013, 312 с.

[6] Шевкунова Е.С. Анализ потребления продуктов питания. *Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета*, 2014, № 101, с. 480–495.

[7] Уварова В.И., Волков Г.О., Евдокимова О.В. Исследование уровня удовлетворения физиологических потребностей населения в продуктах питания. *Маркетинг в России и за рубежом*, 2006, № 1, с. 48–53.

[8] Якимов А.С., Баженов Р.И. Сегментация клиентов с помощью RFM-анализа. *Экономика и менеджмент инновационных технологий*, 2015, № 1, с. 55–61.

[9] Баженов Р.И., Векслер В.А., Гринкруг Л.С. RFM-анализ клиентской базы в прикладном решении 1С: Предприятие 8.3. *Информатизация и связь,* 2014, № 2, с. 51–54.

[10] Голубков Е.П. RFM-анализ: методика и практика применения. *Маркетинг в России и за рубежом,* 2013, № 6, с. 11–24.

[11] Александров В.И. Применение RFM-анализа при разработке таргетированных маркетинговых стратегий в сфере e-commerce. *Маркетинг и маркетинговые исследования,* 2014, № 5, с. 332–339.

[12] Коробов П.Н. *Математическое программирование и моделирование экономических процессов*. СПб., ДНК, 2002, 364 с.

[13] Иванилов Ю.П., Логов Л.В. *Математическое моделирование в экономике.* Москва, Наука, 1979, 304 с.

[14] Печерских И.А., Семенов А.Г. *Математические модели в экономике: учебное пособие.* Кемерово, КемТИПП, 2011, 191 с.

[15] Алесинская Т.В., Сербин В.Д., Катаев А.В. *Учебно-методическое пособие по курсу «Экономико-математические методы и модели. Линейное программирование».* Таганрог, Изд-во ТРТУ, 2001, 79 с.

[16] Кобелев Н.Б. *Практика применения экономико-математических методов и моделей.* Москва, Наука, 2000, 248 с.

[17] Солопахо А.В. *Математика в экономике: учебно-практическое пособие. Ч.1.* Тамбов, Изд-во ТГТУ, 2001, 71 с.

[18] Поттосина С.А., Журавлев В.А. *Экономико-математические модели и методы: учеб. пособие для студ. экон. спец. БГУИР всех форм обуч.* Минск, БГУИР, 2003, 94 с.

[19] Алесинская Т.В. *Учебное пособие по решению задач по курсу «Экономико-математические методы и модели».* Таганрог, Изд-во ТРТУ, 2002, 153 с.

[20] Димитриенко Ю.И., Димитриенко О.Ю. Модель деформируемых кластеров для анализа динамических данных в экономике. *Информационные технологии*, 2010, № 9, с. 43–50.
[21] Димитриенко Ю.И., Димитриенко О.Ю. Кластерно-континуальное моделирование экономических процессов. *Доклады Академии наук*, 2010, т. 435, № 4, с. 466-469.
[22] Димитриенко Ю.И., Димитриенко О.Ю. Кластерно-континуальное моделирование в экономике на основе методов механики многомерных сплошных сред. *Информационные технологии*, 2010, № 8, с. 54–62.
[23] Димитриенко Ю.И., Димитриенко О.Ю. Модель многомерной деформируемой сплошной среды для прогнозирования динамики больших массивов индивидуальных данных. *Математическое моделирование и численные методы*, 2016, № 1(9), с. 105–122.
[24] Dimitrienko Yu.I., Dimitrienko O.Yu. Application of Continuum Mechanics Methods for Economy. *Journal of Physics: Conference Series, 2018,* vol. 1141, no. 012019. DOI: 10.1088/1742-6596/1141/1/012019

Статья поступила в редакцию 24.09.2019

Ссылку на эту статью просим оформлять следующим образом:
Димитриенко Ю.И., Котельникова А.В. Задачи идентификации индивидуальных покупателей на основе анализа больших объемов панельных данных о кассовых чеках. *Математическое моделирование и численные методы*, 2019, № 4, с. 100–116.

Димитриенко Юрий Иванович — д-р физ.-мат. наук, заведующий кафедрой «Вычислительная математика и математическая физика» МГТУ им. Н.Э. Баумана, директор Научно-образовательного центра «Суперкомпьютерное инженерное моделирование и разработка программных комплексов» МГТУ им. Н.Э. Баумана. Автор более 400 научных работ в области механики сплошной среды, вычислительной механики, газодинамики, механики и термомеханики композитов, математического моделирования в науке о материалах, моделирования в экономике. e-mail: dimit.bmstu@gmail.com

Котельникова Александра Васильевна — аспирант кафедры «Вычислительная математика и математическая физика» МГТУ им. Н.Э. Баумана.
e-mail: sasha-dobrynina@yandex.ru

The problems of identifying individual customers based on the analysis of large volumes of panel data on cash receipts

© Yu.I. Dimitrienko, A. V. Kotel'nikova

Bauman Moscow State Technical University, Moscow, 105005, Russia

The problems of identifying individual customers are formulated based on the analysis of large amounts of data on cash receipts in a large supermarket. Models of behavior of various categories of individual customers in the supermarket are developed. A computational algorithm is proposed for solving the problems of identifying individual customers using panel data from cash receipts. The algorithm is universal, since it does not use any personal data about the buyer, but is based on an analysis of only buying activity, calculated on the basis of panel data on cash receipts. The algorithm allows you to identify groups of customers, as well as with a certain probability, an individual customer. As an

example of the application of the developed models and computational algorithms, commodity checks from the X5Retail Group company supermarket chain for a certain period of time were used.

Keywords: *individual customer, identification problems, clustering, cash receipts, big data analysis, panel data, forecasting*

REFERENCES

[1] Barsegyan A.A., Kupriyanov M.S., Holod I.I. Tess M.D. Elizarov S.I. *Analiz dannyh i processov: uchebnoe posobie*, *3-e izd*. [Analysis of data and processes: textbook, 3rd ed.]. Spb., BHV-Peterburg Publ., 2009, 512 p.

[2] Demidova L.A., Kirakovskij V.V., Pyl'kin A.N. *Prinyatie reshenij v usloviyah neopredelennosti* [Making decisions in an uncertain environment]. Moscow, Goryachaya liniya — Telekom, 2012, 288 p.

[3] Demin I.S. *Klasterizaciya kak instrument intellektual'nogo analiza dannyh. Ch. 1: Novye informacionnye tekhnologii v obrazovanii* [Clusterization as a tool for data mining. Part 1: New information technologies in education]. Moscow, 1 S-Pablishing, 2011, pp. 98–103.

[4] Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. *Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya* [Recognition. Mathematical method. Software system. Practical application]. Moscow, Fazis Publ., 2006, 176 p.

[5] Kulaichev A.P. *Metody i sredstva kompleksnogo analiza dannyh: uchebnoe posobie, 4-e izd., pererab. i dop.*[Methods and tools for integrated data analysis: a textbook, 4th ed., rev. and add]. Moscow, FORUM Publ., Infra-M Publ., 2013, 312 p.

[6] Shevkunova E.S. *Politematicheskij setevoj elektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta — Scientific Journal of KubSAU*, 2014, no. 101, p. 480–495.

[7] Uvarova V.I., Volkov G.O., Evdokimova O.V. Marketing v Rossii i za rubezhom — *Journal of Marketing in Russia and Abroad*, 2006, no. 1, pp. 48–53.

[8] Yakimov A.S., Bazhenov R.I. *Ekonomika i menedzhment innovacionnyh tekhnologij — Economics and innovations management*, 2015, no. 1, pp. 55–61.

[9] Bazhenov R.I., Veksler V.A., Grinkrug L.S. Informatizaciya i svyaz' — *Informatization and communication*, 2014, no. 2, pp. 51–54.

[10] Golubkov E.P. Marketing v Rossii i za rubezhom — *Journal of Marketing in Russia and Abroad*, 2013, no. 6, pp. 11–24.

[11] Aleksandrov V.I. *Marketing i marketingovye issledovaniya — Marketing and market research*, 2014, no. 5, pp. 332–339.

[12] Korobov P.N. *Matematicheskoe programmirovanie i modelirovanie ekonomicheskih processov* [Mathematical programming and modeling of economic processes]. SPb., DNK Publ., 2002, 364 p.

[13] Ivanilov Yu.P., Logov LV. *Matematicheskoe modelirovanie v ekonomike* [Mathematical modeling in Economics]. Moscow, Nauka Publ., 1979, 304 p.

[14] Pechersky I. A., Semenov A. G. *Matematicheskie modeli v ekonomike: uchebnoe posobie* [Mathematical models in Economics: textbook]. Kemerovo, KemTIPP Publ., 2011, 191 p.

[15] Alesinskaya T. V., Serbin V. D., Kataev A.V. *Uchebno-metodicheskoe posobie po kursu «Ekonomiko-matematicheskie metody i modeli. Linejnoe programmirovanie»* [Educational and methodical manual for the course "

Economic and mathematical methods and models. Linear programming"]. Taganrog, TRTU Publ., 2001, 79 p.

[16] Kobelev N.B. *Praktika primeneniya ekonomiko-matematicheskih metodov i modelej* [Practice of applying economic and mathematical methods and models]. Moscow, Nauka Publ., 2000, 248 p.

[17] Solopaho A.V. *Matematika v ekonomike: uchebno-prakticheskoe posobie. Ch.1* [Mathematics in Economics: educational and practical guide. Part 1]. Tambov, TSTU Publ., 2001, 71 p.

[18] Pottosina S.A., Zhuravlev V.A. *Ekonomiko-matematicheskie modeli i metody: ucheb. posobie dlya stud. ekon. spec. BGUIR vsekh form obuch.*[Economic and mathematical models and methods: textbook. a manual for students. ekon. special BGUIR of all forms of education]. Minsk, BGUIR Publ., 2003, 94 p.

[19] Alesinskaya T.V. *Uchebnoe posobie po resheniyu zadach po kursu «Ekonomiko-matematicheskie metody i modeli*» [Textbook for solving problems in the course "Economic and mathematical methods and models"]. Taganrog, TRTU Publ., 2002, 153 p.

[20] Dimitrienko Yu.I., Dimitrienko O.Yu. *Informacionnye tekhnologii — Information Technologies*, 2010, no. 9, pp. 43–50.

[21] Dimitrienko Yu.I., Dimitrienko O.Yu. Doklady *Akademii nauk — Doklady Akademii Nauk*, 2010, vol. 435, no. 4, pp. 466-469.

[22] Dimitrienko Yu.I., Dimitrienko O.Yu. *Informacionnye tekhnologii — Information Technologies*, 2010, no. 8, pp. 54–62.

[23] Dimitrienko Yu.I., Dimitrienko O.Yu. *Matematicheskoe modelirovanie i chislennye metody — Mathematical modeling and Computational Methods*, 2016, no. 1(9), pp. 105–122.

[24] Dimitrienko Yu.I., Dimitrienko O.Yu. Application of Continuum Mechanics Methods for Economy. *Journal of Physics: Conference Series, 2018,* vol. 1141, no. 012019. DOI: 10.1088/1742-6596/1141/1/012019

Dimitrienko Yu. I., Professor, Head of the Department of Computational Mathematics and Mathematical Physics, Director of Scientific-Educational Center of Supercomputer Engineering Modeling and Program Software Development, Bauman Moscow State Technical University. Member of the Russian Academy of Engineering Science. Author of over 400 research publications in the field of computational mechanics, gasdynamics, thermomechanics of composite materials, mathematical simulations in material science, modeling in economy. e-mail: dimit.bmtstu@gmail.com

Kotel'nikova A.V., postgraduate of the Computational Mathematics and Mathematical Physics Department at Bauman Moscow State Technical University.
e-mail: sasha-dobrynina@yandex.ru