ПРОГНОЗ ПОВЕДЕНИЯ КЛИЕНТОВ СУПЕРМАРКЕТОВ С ПОМОЩЬЮ ВЕСОВЫХ СХЕМ ОЦЕНОК ВЕРОЯТНОСТЕЙ И ПЛОТНОСТЕЙ[1]

А.Г. Дьяконов,
доктор физико-математических наук, профессор кафедры математических методов прогнозирования, факультет вычислительной математики и кибернетики, Московский государственный университет им. М.В. Ломоносова, старший научный сотрудник, Вычислительный центр им. А.А. Дородницына Российской академии наук

Адрес: 119991, Москва, ГСП-1, Ленинские горы, МГУ им. М.В. Ломоносова, 1, стр. 52
E-mail: djakonov@mail.ru

Рассмотрены две задачи, связанные с поведением клиентов сети супермаркетов: прогнозирование даты следующего визита каждого клиента и суммы его покупок. Первая задача сведена к задаче оценки вероятностей визитов, вторая – к задаче восстановления плотностей распределений сумм покупок каждого пользователя. Для решения указанных задач предложено использовать взвешенные схемы: каждой точке выборки ставится в соответствие вещественное неотрицательное число (вес). Веса позволяют учитывать дополнительную информацию, например устаревание данных (точки соответствующие старым данным имеют меньшие веса). В работе рассмотрено несколько весовых схем (способов приписывания весов точкам выборки), произведена их настройка (оптимизация качества оценки вероятности или плотности по параметрам весовой схемы). Показано, что использование весовых схем не приводит к переобучению, т.е. настройка весов на обучении не понижает качество на независимой контрольной выборке. Показана возможность использования ансамблирования для повышения качества решения рассмотренных задач, т.е. построения нескольких алгоритмов и составления их линейной комбинации. Все эксперименты произведены на реальных данных крупного Международного конкурса по разработке алгоритмов анализа данных. Специфика данных (отсутствие праздников на финальном временном отрезке статистики) позволила при решении указанных задач сосредоточиться исключительно на статистических методах решения. Кроме того, рассмотрены вопросы построения алгоритмов, которые одновременно решают обе задачи: прогнозирования даты следующего визита и суммы покупок. Показано, что не всегда их можно решать независимо. Предложен метод оптимизации функционала, который оценивает решение обеих задач.

Ключевые слова: прогнозирование, оценивание вероятности, восстановление плотности, непараметрические методы, прикладные задачи.

[1] Работа поддержана грантами РФФИ № 12-07-00187, №14-07-00965.

Введение

Пусть есть конечное множество клиентов сети супермаркетов. Для построения и исследования алгоритмов оно предполагается фиксированным. Для каждого клиента известны все даты его визитов и суммы покупок в каждую из этих дат. Статистика дана за последние d недель. Обычно подобная статистика собирается по покупкам с использованием скидочных карт (что позволяет идентифицировать клиентов), поэтому рассматриваемое множество не просто множество людей, посещающих магазины сети, а людей, которые достаточно часто посещают эти магазины. Необходимо предсказать дату следующего визита и сумму покупок. Что понимается под точностью прогноза, будет оговорено далее.

Сначала введём необходимый формализм. Статистику посещений конкретного клиента можно записать в виде матрицы с неотрицательными элементами

$$S = \|s_{ij}\|_{d\times 7}, \qquad (1)$$

где s_{ij} – сумма покупок клиента на j-й день i недель назад. Например, если сегодня конец воскресенья, то среда этой недели соответствует элементу s_{13}, а пятница прошлой – s_{25}. Считаем, что визит был тогда и только тогда, когда сумма покупок положительна. Хотя теоретически допустима ситуация «клиент пришёл, но ничего не купил», практически она не отображается в статистике (которая, как уже говорилось, формируется по чекам покупок с дисконтными картами). В дальнейшем будем рассматривать каждого клиента по отдельности (поэтому обозначение (1) не содержит номер клиента).

Отметим, что данные, на которых тестировались методы, охватывали период чуть больше года. Последние месяцы этого периода не включали праздники (Новый год, 8 марта и т.п.) и особые дни (например, 1 сентября в РФ). Поэтому считаем статистику однородной, без влияния сезонных факторов. С одной стороны, это существенно упрощает задачу. С другой, делает возможным тестирование различных методов оценки вероятностей и плотностей распределений.

Все графики и результаты экспериментов представлены для данных компании dunnhumby, которые использовались для проведения Международного конкурса [1]. Автор работы стал победителем этого соревнования среди 279 участников. Здесь представлены результаты последующих исследований, кроме того, все выводы согласуются с экспериментами, которые были сделаны также и на данных российских Интернет-магазинов.

В работе некоторые модификации алгоритмов дают улучшения на доли процентов. Отметим, что это действительно улучшения (качество увеличивается и на независимом контроле и на других данных). Кроме того, в современных бизнес-задачах даже незначительные улучшения могут дать существенный доход, применение же нескольких модификаций увеличивает качество на проценты. В режиме соревнования [1] именно эти модификации позволили построить лучший алгоритм.

1. Решение задачи прогноза даты следующего визита

1.1. Оценивание вероятностей, пересчёт, весовые схемы

Для каждого клиента необходимо предсказать день его следующего визита: $j \in \{1, 2, ...\}$ (1 соответствует завтрашнему дню, 2 – послезавтрашнему и т.д.). Прогноз считается верным, если мы угадываем этот день, т.е. клиент придёт в j-й день и не придёт в 1, 2, ..., (j -1)-й день. Далее при оценке методов будем указывать процент верных прогнозов (по всем клиентам): из данных удаляется информация о последней неделе, метод прогнозирует день визита каждого клиента на этой удалённой неделе, процент таких верных прогнозов и считается качеством метода.

Имеет смысл перейти от матрицы сумм покупок $S = \|s_{ij}\|_{d\times 7}$ к бинарной матрице визитов $V = \|v_{ij}\|_{d\times 7}$:

$$v_{ij} = 1 \Leftrightarrow s_{ij} > 0.$$

Предложим вероятностную модель поведения клиента: в j-й день недели он с вероятностью p_j посещает магазин. Напомним, что первый день недели – это день, которому соответствует $j = 1$ при формировании матрицы (1) (прогнозирование происходит в конце 7-го дня, сразу перед началом новой недели). В такой вероятностной модели вполне естественно вычислить вероятности первых визитов в j-е дни, что делается по следующей «формуле пересчёта»:

$$\tilde{p}_j^1 = p_j \prod_{r=1}^{j-1} (1 - p_r), \qquad (2)$$

(считаем, что $\tilde{p}_1^1 = p_1$) и выбрать из них максимальное значение:

$$j = \underset{r\in\{1,2,\ldots,7\}}{argmax}\, \tilde{p}_r^1 .$$

В этой вероятностной модели неявно делаются следующие предположения:

1. Поведения разных клиентов сети супермаркетов независимы (поэтому ответ для конкретного клиента зависит только от его статистики посещений).

2. Каждый клиент обязательно посетит магазин в течение следующей недели.

Отметим, что хотя первое предположение вряд ли верно, пока не удалось предложить методы, использующие зависимости в поведении разных клиентов, которые превосходили бы по качеству описанные в данной работе.

Второе предположение легко обойти, «увеличив неделю», например, до 14 дней. В этом случае считаем, что клиент обязательно посетит магазин в течение ближайших 14 дней. Как показывают эксперименты, качество алгоритма не меняется при таком «увеличении недели». Важно понимать, что очень много зависит от свойств данных, на которых предполагается использовать метод. Далее приводятся результаты экспериментов на реальных данных [1], для которых алгоритм «клиент придёт завтра» уже даёт более 25% верных прогнозов (*см. рис. 1*). Отметим также, что в соответствии с этими данными за неделю магазины посещают около 85% клиентов (держателей дисконтных карт).

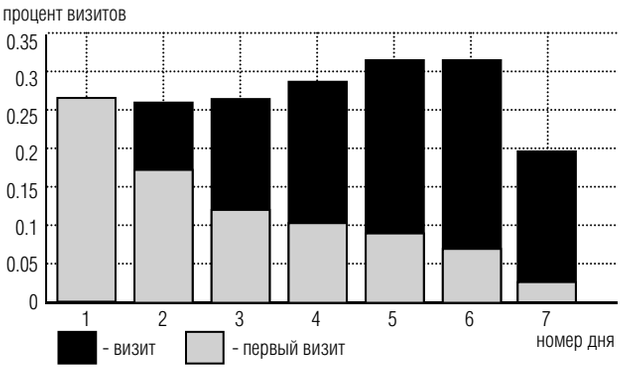

Рис. 1. Процент визитов и первых визитов на неделе в зависимости от дня недели

Простейший (и самый распространенный на практике) способ оценивания вероятности – через частоту, в данном случае, вероятности p_j можно оценить через частоты визитов за последние d недель:

$$p_j = \frac{1}{d}\sum_{i=1}^{d} v_{ij}. \qquad (3)$$

Однако необходимо учитывать, что информация последних недель ценнее информации d недель назад, поэтому можно рассмотреть взвешенную схему оценки вероятности:

$$p_j = \sum_{i=1}^{d} w_i v_{ij},$$

$$w_1 \geq w_2 \geq \ldots \geq w_d \geq 0, \sum_{i=1}^{d} w_i = 1.$$

Подобные весовые схемы применяются в анализе данных, например во взвешенном методе ближайших соседей [2]. Часто достаточно эффективный способ задания весов следующий:

$$w_i^N = \left(\frac{d-i+1}{d}\right)^{\delta}, i \in \{1,2,\ldots,d\}, \qquad (4)$$

$$w_i = \frac{w_i^N}{\sum_{i=1}^{d} w_i^N}, i \in \{1,2,\ldots,d\}. \qquad (5)$$

Здесь выбор конкретной весовой схемы сводится к выбору параметра $\delta \in [0, +\infty)$. При этом, значение $\delta = 0$ соответствует равным весам, т.е. вычислению вероятностей по формуле (3).

В дальнейшем при задании весов будем как в (4) использовать верхний индекс N, подразумевая, что далее необходима нормировка по сумме (5), которую явно указывать не будем.

Как видно из *рис. 2* (сплошная линия) применение схемы весов позволяет немного повысить качество прогноза: простой метод по формулам (3), (2) даёт точность 35.9% , взвешенный – около 36.36%. Перестановка и прямой метод (качество которых также изображено на *рис. 2*) будут введены чуть ниже (в конце этого раздела и в разделе 1.2).

Как и во взвешенном методе ближайших соседей [2], можно выбирать ещё параметр «число ненулевых весов» k:

$$w_i^N = \begin{cases} \left(\frac{d-i+1}{d}\right)^{\delta}, & i \in \{1,2,\ldots,k\}, \\ 0, & i \in \{k+1,\ldots,d\}. \end{cases}$$

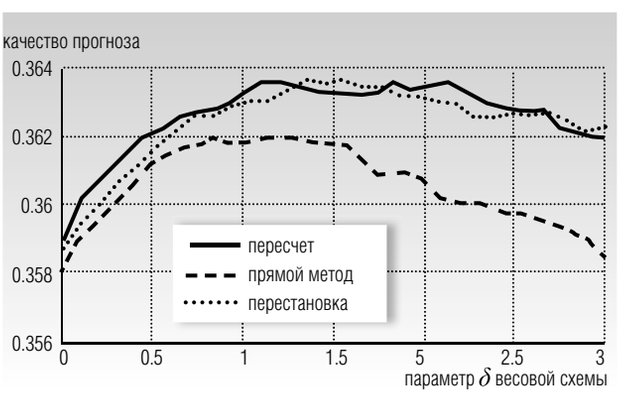

Рис. 2. Зависимость качества прогноза от степени δ

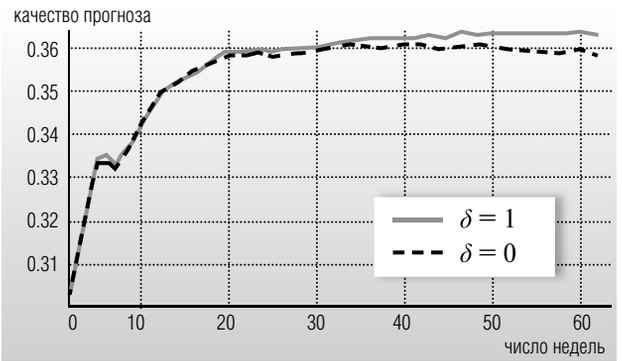

Рис. 3. Зависимость качества прогноза от числа учитываемых недель

Но, как видно на *рис. 3*, здесь работает принцип «чем больше данных, тем лучше». Кстати, простой алгоритм «будет как на прошлой неделе» ($k=1$) показывает качество 29.3% (лучше константного $j=1$, но существенно хуже рассматриваемых).

Также на практике иногда удаляют куски данных «без информации». В данном случае – нулевые строки из матрицы $V=\|v_{ij}\|_{d\times 7}$, чтобы не учитывались недели, в которые не было визитов. Для удобства (это упрощает программирование во многих пакетах, например Matlab), они остаются в матрице, просто «съезжают вниз», т.е. если

$$I=\{i_1,\ldots,i_r\}=\{i \mid v_{i1}+\ldots+v_{i7}>0\}, 1\le i_1<\ldots<i_r\le d,$$

то матрица $V=\|v_{ij}\|_{d\times 7}$ заменяется матрицей

$$\begin{Vmatrix} v_{i_1,1} & \cdots & v_{i_1,7} \\ \cdots & \cdots & \cdots \\ v_{i_r,1} & \cdots & v_{i_r,7} \\ 0 & \cdots & 0 \\ \cdots & \cdots & \cdots \end{Vmatrix}$$

Как видно из *рис. 2* (чёрные точки), такая перестановка недель не сильно увеличивает качество (36.37%), но делает функцию качества от параметра $\delta\in[0,+\infty)$ «почти унимодальной».

Вообще, весовые схемы можно выбирать поразному. Например,

$$w_i^N=\lambda^i,\ i\in\{1,2,\ldots,d\},\lambda\in(0,1], \quad (6)$$

или

$$w_i^N=\frac{1}{i^\gamma}, i\in\{1,2,\ldots,d\},\ \gamma\in[0,+\infty)\cdot \quad (7)$$

На *рис. 4* показаны распределения весов (после нормировки по сумме (5)) при первой схеме (4) для различных $\delta\in[0, 3]$ (слева), при второй (6) для различных $\lambda\in[0.9, 1]$ (в центре) и при третьей (7) для различных $\gamma\in[0, 1]$ (справа). В каждом случае отрезок возможных значений параметра делился на равные части 30-ю точками и строились графики для 30-ти различных значений параметра. Зависимости качества от параметров при различных нормировках похожи, ср., например, *рис. 2* и *рис. 5*.

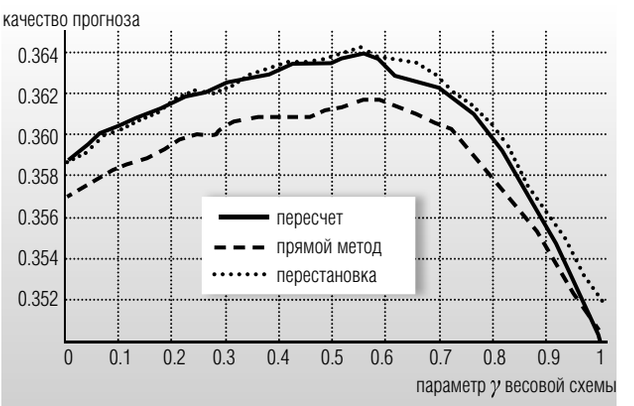

Рис. 5. Зависимость качества прогноза от $\gamma\in[0, 1]$

1.2. Прямое оценивание вероятностей

Оценить вероятности первых визитов можно и непосредственно (без пересчёта по оценкам вероятностей визитов):

$$\tilde{p}_j^2=\frac{1}{d}\,|\,\{i\in\{1,2,\ldots,d\}: v_{i1}=\ldots=v_{i,\,j-1}=0,\ v_{ij}=1\}|, \quad (8)$$

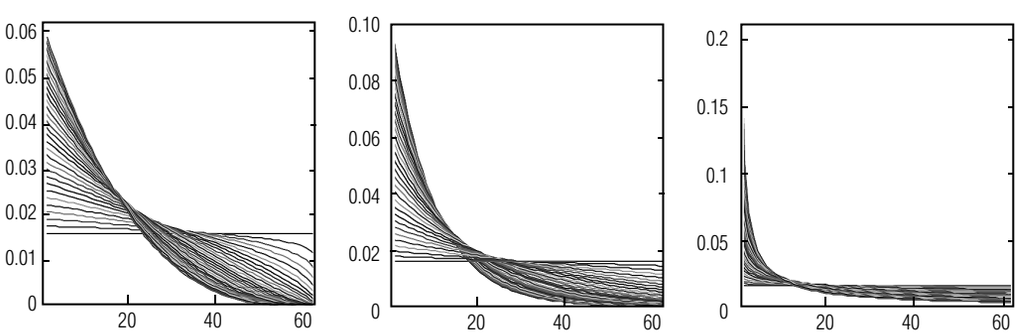

Рис. 4. Три разные весовые схемы: вес недели в зависимости от её номера

т.е. для каждого дня недели посчитать долю недель, в которые в этот день был первый визит. Такой метод кажется даже «более естественным». Если в матрице V в каждой строке оставить лишь первый единичный элемент (если он есть), а остальные занулить, т.е. перейти к матице первых визитов $V' = \|v'_{ij}\|_{d\times 7}$, то для оценки вероятности можно использовать следующую взвешенную схему:

$$\tilde{p}_j^2 = \sum_{i=1}^{d} w_i v'_{ij}. \qquad (9)$$

Равные веса соответствуют формуле (8). Отметим, что это окончательная формула (пересчёта вероятностей (2) не требуется).

Из *рис. 2, 5* (прерывистая линия) видно, что качество прямого метода чуть ниже, чем описанного ранее метода с пересчётом вероятностей.

1.3. Ансамблирование

Часто качество прогнозирования повышается при «ансамблировании» нескольких методов, т.е. построении алгоритма на базе разных методов, тогда недостатки одного метода, компенсируются достоинствами других. На принципе ансамблирования построены как многие современные эффективные алгоритмы машинного обучения (бэггинг, бустинг и т.д.), так и некоторые теории анализа и построения алгоритмов (например, алгебраический подход, комитетный подход и т.д.) [3].

Выше предложены две оценки вероятности первого визита (2) и (8). Первый способ ансамблирования – «стандартный ансамбль» – взять их выпуклую комбинацию:

$$\tilde{p}_j = \alpha\tilde{p}_j^1 + (1-\alpha)\tilde{p}_j^2, \qquad (10)$$

где $\alpha \in [0, 1]$ – параметр для настройки.

Предложим другой способ ансамблирования: будем брать не выпуклую комбинацию $\tilde{p}_j^1$ и $\tilde{p}_j^2$, а выпуклую комбинацию p_j и $\tilde{p}_j^2$:

$$\alpha p_j + (1-\alpha)\tilde{p}_j^2 = \alpha\sum_{i=1}^{d} w_i v_{ij} + (1-\alpha)\sum_{i=1}^{d} w_i v'_{ij} = \sum_{i=1}^{d} w_i(\alpha v_{ij} + (1-\alpha)v'_{ij}) \qquad (11)$$

(после этого требуется пересчёт вида (2) для определения вероятностей первых визитов). Такой метод не имеет строгого теоретического обоснования (хотя в [4] представлены некоторые аргументы в его пользу), но очень неплохо показал себя на практике. Его можно интерпретировать как более сложное задание весов, см. (11). На *рис. 6* видно, что предложенный

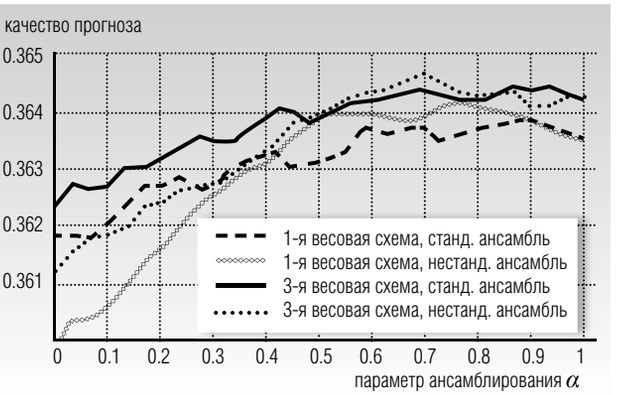

Рис. 6. Качество ансамблирования от параметра $\alpha \in [0, 1]$

«нестандартный ансамбль» лучше «стандартного» и позволяет повысить качество с 36.35% до 36.41% в случае первой весовой схемы (4) и с 36.42% до 36.46% – в случае использования третьей (7).

1.4. Проблема переобучения

Переобучением в анализе данных называется эффект, при котором на новых данных алгоритм работает существенно хуже, чем на исходных (на основе которых он был построен) [5]. Как правило, переобучение связывают с излишней сложностью модели алгоритмов. Для проверки отсутствия переобучения используют отложенный контроль: часть данных не используют при настройке параметров, затем на этой части (отложенной выборке) проверяют качество алгоритма.

Все приводимые графики получены для фиксированной выборки. Естественно, встаёт вопрос о надёжности алгоритма, т.е. как он будет работать на новых данных. Оказывается, что предложенные взвешенные схемы с оптимальными параметрами практически не склонны к переобучению. На *рис. 7* представлена столбцовая диаграмма с результатами экспериментов на обучающей выборке и отложенном контроле (в качестве контроля использовалась последняя неделя, за которую есть статистика визитов клиентов, она не использовалась для настройки параметров). Показано качество следующих алгоритмов:

1) константный («клиент придёт на следующий день»),
2) визит клиента как на прошлой неделе,
3) вероятности (3) оценены по последним 5 неделям,
4) вероятности оценены по всем неделям,
5) оптимальные значения весов (7),
6) оптимальное нестандартное ансамблирование (11).

Несмотря на кажущееся усложнение алгоритма (ввод весов, ансамблирование), не только повы-

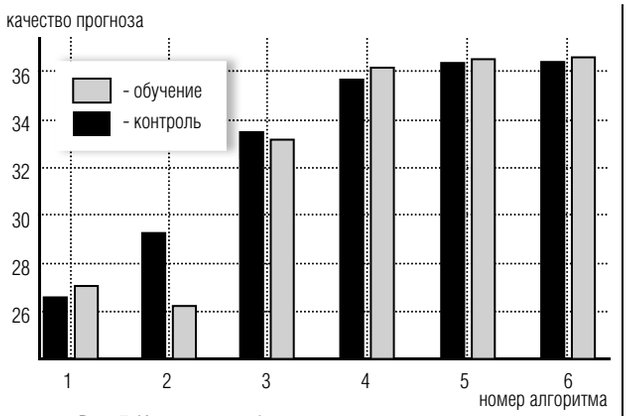

Рис. 7. Качество на обучении и отложенном контроле для шести алгоритмов

шается качество, но и уменьшается разница между обучением и контролем. Поэтому далее качество на отложенном контроле не будет приводиться (оно почти такое же, как на обучении).

2. Решение задачи прогноза суммы покупок

В день визита клиента в супермаркет необходимо предсказать его траты с точностью до $\varepsilon = 10$ у.е., т.е. для t-го клиента ответ a^t алгоритма прогнозирования считается верным, если $|s^t - a^t| \le \varepsilon$, где s^t – истинные траты клиента. Качество на всей выборке оценивается как процент верных ответов:

$$\frac{1}{n}\sum_{t=1}^{n}\begin{cases}1, & |s^t - a^t| \le \varepsilon,\\ 0, & |s^t - a^t| > \varepsilon,\end{cases}$$

где n – число клиентов (для которых сделан прогноз). По нашей договорённости дальше индекс клиента t будем опускать, решая задачу прогнозирования независимо для каждого клиента.

Эту задачу можно решать как обычную задачу восстановления плотности. Пусть клиент делал покупки на суммы $s_1, \ldots, s_m$. Будем считать, что это реализации некоторой случайной величины. Если оценить плотность её распределения, то разумно в качестве прогноза выдать значение s, в котором плотность максимальна. Есть три подхода к оцениванию плотности [6]:

1) параметрический (когда известно распределение с точностью до параметров),

2) смеси распределений (когда известно, что плотность представляется в виде выпуклой комбинации плотностей, известных с точностью до параметров),

3) непараметрический (когда не известен вид распределения).

Для каждого клиента траты достаточно сильно отличаются (*см. рис. 8*), кроме того, часто от дня недели зависят суммы покупок и разброс этих сумм (*рис. 9*). Для решения задачи был выбран достаточно универсальный непараметрический подход, поскольку:

1) вид распределения априорно не известен,

2) проще настраивать плотность, а функционал качества автоматически определяет ядерную функцию (см. далее),

3) подход допускает обобщения с использованием весовых схем (и, соответственно, учёт времени покупок).

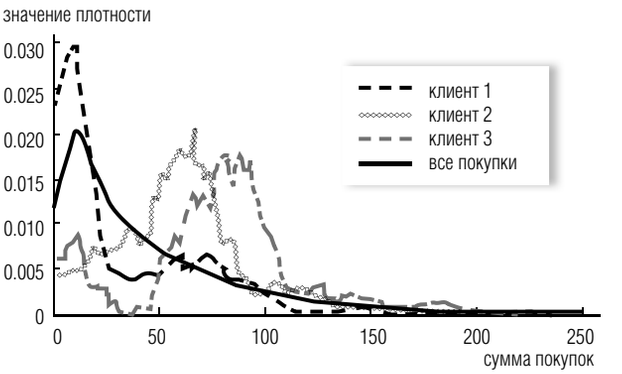

Рис. 8. Плотности распределения покупок

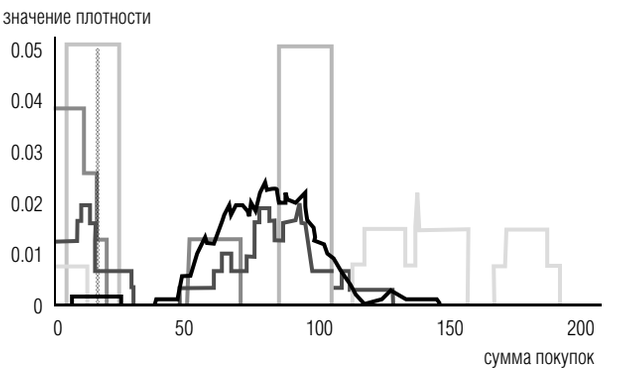

Рис. 9. Плотности покупок одного пользователя в разные дни недели

Для оценки плотности непараметрическим методом Розенблатта–Парзена [7, 8] пользуемся формулой

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} K(s_i - x), \qquad (12)$$

где функция $K(x)$ – ядерная (ядро), т.е. симметричная неотрицательная вещественная функция, монотонно не возрастающая на положительной полуоси, интегрируемая, причём

$$\int_{-\infty}^{+\infty} K(x)\,dx = 1.$$

Эти требования обеспечат, чтобы выражение (12) имело смысл плотности.

Функционал качества диктует выбор ядра:

$$K(s-x)=\begin{cases}1/2\varepsilon, & |s-x|\le\varepsilon,\\ 0, & |s-x|>\varepsilon.\end{cases}$$

Это становится ясно из различных модельных примеров. Например, если $s_1=\ldots=s_m=s$, т.е. клиент всё время делает покупки на одну и ту же сумму, то алгоритм может в качестве прогноза выдавать любое значение из отрезка $[s-\varepsilon, s+\varepsilon]$. Если же, например, клиент совершает покупки только на суммы из двухэлементного множества $\{s-\varepsilon, s+\varepsilon\}$, то алгоритм должен прогнозировать значение s (только так он «угадывает» в обоих случаях). При описанном выборе ядра выражение (12) в первом случае достигает максимального значения на отрезке $[s-\varepsilon, s+\varepsilon]$, а во втором – в точке s.

Отметим, что на *рис. 8 – 9* изображены плотности, полученные по формуле (12). На *рис. 8* изображена плотность для всех покупок (всех пользователей) и для трёх пользователей, совершивших 182, 70 и 80 покупок соответственно. Для последнего на *рис. 9* показана более подробная информация по дням недели.

Формула (12) легко обобщается на случай использования весов:

$$f(x)=\sum_{i=1}^{m} w_i K(s_i-x), \qquad (13)$$

поэтому оценка плотности однозначно определяется весами w_i, $i\in\{1, 2, \ldots, m\}$. Веса имеет смысл использовать, поскольку ценность информации о покупке на сумму s_i зависит от того, в какой день недели была сделана покупка и как давно. Далее тот факт, что i-й покупке на сумму s_i соответствует вес w_i будем обозначать так:

$$s_i \leftrightarrow w_i$$

(это не вызовет путаницы). При этом если сумма всех весов w_i отлична от единицы, то считаем, что перед использованием формулы (13) необходимо сделать нормировку: поделить каждый вес на эту сумму.

Рассмотрим следующий выбор весов. Пусть $s_1, \ldots, s_m$ – все покупки пользователя, упорядоченные по дате от самой последней до самой первой, $s'_1,\ldots,s'_{m'}$ – покупки, сделанные в день недели, на который делается прогноз, упорядоченные аналогично. Плотность по формуле (13) будем восстанавливать для расширенного набора $s'_1,\ldots,s'_{m'},s_1,\ldots,s_m$. Некоторые покупки здесь учитываются дважды. Это соответствует тому, что их вес будет равен сумме двух весов (для первого учёта и для второго). Пусть

$$s'_i \leftrightarrow \beta\frac{(m'-i+1)^{\rho'}}{\sum_{j=1}^{m'} j^{\rho'}}, i\in\{1,2,\ldots,m'\} \qquad (14)$$

$$s_i \leftrightarrow (1-\beta)\frac{(m-i+1)^{\rho}}{\sum_{j=1}^{m} j^{\rho}}, i\in\{1,2,\ldots,m\}. \qquad (15)$$

Параметр β определяет вклад каждого из двух поднаборов, степени ρ, ρ' – весовые схемы для каждого из поднаборов. Отметим, что здесь происходит целых три нормировки: отдельно для каждого из поднаборов (14), (15), а затем общая нормировка перед применением формулы (13).

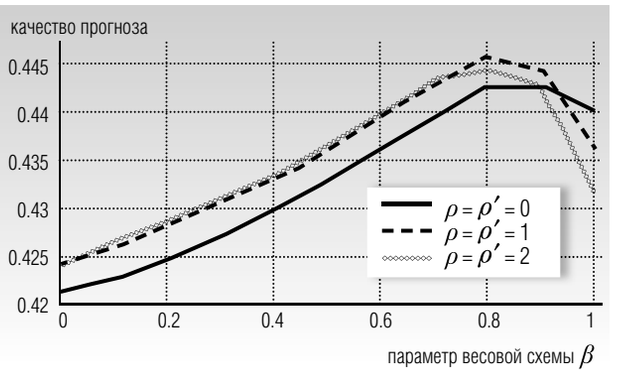

Рис. 10. Качество прогноза суммы покупок от параметра β

На *рис. 10* показано, что в случае, когда мы не используем весовую схему ($\beta=0$, $\rho=0$) качество прогноза – 42.12%, использование весов повышает качество до 44.55%. При настройке параметров хорошо работает метод покоординатного спуска, поскольку, как видно на графиках, функция качества «достаточно выпуклая». При $\rho=0.75$, $\rho'=1.25$, $\beta=0.8$ качество прогноза достигает 44.68%.

Конечно, очень важно настраивать параметры на «правильной выборке». Можно для всех клиентов взять первый визит на последней неделе и по этим данным производить настройку, а можно сначала запустить алгоритм прогноза дня первого визита и настройку производить только по угаданным визитам. Дальше мы на этом остановимся подробнее, пока отметим, что в данном случае (*см. рис. 11*) выборка не влияет на значения оптимальных параметров, хотя качество прогноза суммы покупок на угадываемых днях выше: более 48.5% (т.е. в ожидаемые дни клиент ведёт себя более предсказуемо).

Вообще, в этой задаче можно придумать много разных весовых схем. К сожалению, для нахождения лучшей их нельзя «перебрать». В некоторых

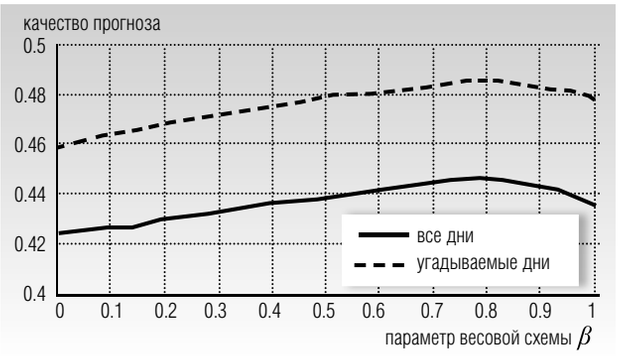

Рис. 11. Качество прогноза суммы покупок от параметра β при $\rho = 0.7$, $\rho' = 1.6$

прикладных задачах применялись такие «экзотические» схемы: формировался набор $s''_1,\ldots,s''_{m''}$ из первых $\tilde{m} = min(m',\omega')$ элементов набора $s'_1,\ldots,s'_{m'}$ и первых $min(m,\omega+\lfloor\sigma\tilde{m}\rfloor)$ элементов набора s_1, ..., s_m, использовалась степенная схема весов

$$s''_i \;\leftrightarrow\; (m''-i+1)^{\rho}, i\in\{1,2,\ldots,m''\}.$$

Здесь $\omega',\omega,\sigma,\rho$ – параметры для настройки.

3. Предсказание поведения клиента

Предположим теперь, что для каждого клиента одновременно надо решить две задачи: предсказать дату визита и сумму покупок. Причём ответ считается верным, если верно решены обе задачи: правильно предсказан день, а ошибка прогноза суммы покупок не больше ε.

Сомнительным выглядит «наивный» метод решения: предсказать день, а потом в этот день предсказать сумму. Действительно, в идеальном случае (когда наши оценки вероятностей точные) вероятность правильности нашего ответа j в первой задаче это и есть максимальная вероятность $\tilde{p}_j$ среди $\tilde{p}_1,\ldots,\tilde{p}_7$ (и именно её мы максимизируем). Пусть теперь q_j – вероятность правильности ответа во второй задаче (прогнозировании трат в j-й день). Тогда необходимо максимизировать $\tilde{p}_j q_j$ по j и q_j, а не последовательно $\tilde{p}_j$ по j и $\tilde{p}_j q_j$ по q_j как в «наивном» методе.

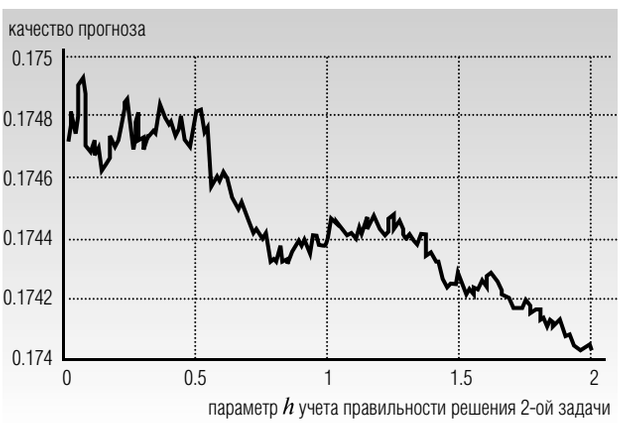

Рис. 12. Качество предсказания поведения в зависимости от параметра h

Предложенный выше метод решения второй задачи позволяет оценить свою надёжность: максимальное значение плотности $f(x)$ это и есть q_j – оценка вероятности правильности ответа.

На *рис. 12* показано качество предсказания поведения клиента (т.е. дня первого визита и суммы трат в этот день) в зависимости от параметра h при максимизации

$$\tilde{p}_j(q_j+h)\to \max_j . \qquad (16)$$

При выборе очень большого h получается наивный метод, в этом случае качество равно 17.38, при малых $h\in[0, 0.5]$ качество около 17.48. Значение $h = 0$ соответствует методу, описанному выше. Таким образом, это параметр регулирует переход от наивного метода к «продвинутому».

Отметим, что, несмотря на простоту метода (16), в соревновании [1] все участники, кроме победителя, использовали наивный метод.

4. Заключение

В данной работе рассмотрена конкретная задача: прогнозирование следующего визита клиента и суммы покупок. Аналогичные задачи возникают практически везде, где клиенты пользуются некоторыми услугами. Например, для репозитория электронных материалов (видео, книги, софт и т.п.): для каждого пользователя предсказать следующий день прихода на сайт и объём скачанных материалов. Или для потребителей Интернет-трафика: предсказать время следующего потребления и объёмы.

Аналогичные эксперименты были проведены для одного из российских Интернет-магазинов. В качестве основной выборки были взяты постоянные клиенты (которые делают заказы не реже чем 1 раз в N дней), поскольку основная часть клиентов (более 90% зарегистрированных на сайте) делает заказы изредка (сделали лишь 1 заказ, совершают заказы только перед крупными праздниками и т.п.) Вид всех графиков и даже оптимальные значения некоторых параметров для этих данных такие же, как и для данных [1]. Отличаются лишь показатели качества.

Методы, рассмотренные в работе, зависят от большого числа параметров. Итоговый метод ре-

шения задачи прогнозирования поведения клиента зависит от следующих параметров:

1) параметр весов γ при оценке вероятностей визитов/первых визитов, см. (7),

2) параметр ансамблирования α, см. (11),

3) первый параметр весовой схемы ρ при оценке суммы покупок, см. (15),

4) второй параметр весовой схемы ρ' при оценке суммы покупок, см. (14),

5) третий параметр весовой схемы β при оценке суммы покупок, см. (14) – (15),

6) параметр h учёта правильности прогноза трат, см. (16).

Естественно, оптимизация по всем этим параметрам – отдельная большая задача. На практике она делалась покоординатным спуском (последовательно фиксировались все параметры, кроме одного, по которому проводилась оптимизация). В работе мы привели графики «срезов качества»: качества прогнозов от значения одного из параметров при фиксированных остальных. Это сделано для того, чтобы показать вид функционала качества: практически все графики изображают унимодальные функции, часто «достаточно гладкие».

Автор выражает благодарность анонимному рецензенту за ценные замечания, которые помогли существенно улучшить статью. ■

Литература

1. Международный конкурс Dunnhumby's Shopper Challenge / Kaggle [Электронный ресурс]: http://www.kaggle.com/c/dunnhumbychallenge (дата обращения 24.11.2013)
2. Воронцов К.В. Метрические алгоритмы классификации // Электронные лекции [Электронный ресурс]: www.ccas.ru/voron/download/MetricAlgs.pdf (дата обращения 24.11.2013)
3. Воронцов К.В. Алгоритмические композиции // Электронные лекции [Электронный ресурс]: http://www.ccas.ru/voron/download/Composition.pdf (дата обращения 24.11.2013)
4. Дьяконов А.Г. Решение задач анализа данных, основанное на линейной комбинации деформаций // Машинное обучение и анализ данных. 2013. Т. 1. № 5. С. 543–554.
5. Воронцов К.В. Комбинаторная теория надёжности обучения по прецедентам. Дис. ... д-ра физ.-мат. наук. М., 2010. 271 с.
6. Расин Д. Непараметрическая эконометрика: вводный курс // Квантиль. 2008. №4. С. 7-56.
7. Parzen E. On Estimation of a Probability Density Function and Mode // Annals of Mathematical Statistics. 1962. No. 33. P. 1065–1076.
8. Rosenblatt M. Remarks on Some Nonparametric Estimates of a Density Function // Annals of Mathematical Statistics. 1956. No. 27. P. 832-837.

SUPERMARKETS CLIENTS BEHAVIOUR FORECASTING BY WEIGHTED METHODS OF PROBABILITY AND DENSITY ESTIMATIONS

Alexander D'YAKONOV
Professor, Department of Mathematical Methods of Forecasting, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University;
Senior Researcher, Dorodnitsyn Computing Center, Russian Academy of Sciences
Address: 1, build. 52, Lomonosov Moscow State University, Leninskie Gory, GSP-1, Moscow, 119991, Russian Federation
E-mail: djakonov@mail.ru

We consider two tasks in describing a supermarkets clients' behavior: prediction of a client's next visit date and prediction of his/her spends. The first problem is equal to estimating visit probability, and the second – to estimating density for visitor spends. To solve these problems, we propose using weighed methods: real non-negative value (weight) is assigned to every event. Weights allow considering additional information, for example history (earlier visits have smaller weights). We consider several weighted schemes (methods of assigning weights to events) and weights optimization (performance optimization by changing weight parameters). The paper shows that weighted methods don't lead to overfitting, i.e. learning on a training set doesn't decrease performance on an independent test set. We can see, that assemblers of different methods can increase performance (we consider linear combination of probabilities estimated by different methods). All experiments are made on real data of large International competition on data mining. The last span of statistics does not contain holidays, which allows concentrating only on statistical methods of problems solving while solving these tasks. Besides, we also considered construction of algorithm to solve the problems (next visit date and spends prediction) simultaneously. It can be seen that the problems not always can be solved independently. We propose a function to estimate solutions of both problems and optimization method for this function.

Key words: forecasting, probability estimation, density estimation, non-parametric methods, applied problems.

References

1. Kaggle (2011) Dunnhumby's Shopper Challenge (electronic resource). Available at: http://www.kaggle.com/c/dunnhumbychallenge (accessed 24 November 2013).
2. Vorontsov K.V. Metricheskie algoritmy klassifikacii [Metrical Algorithms of Classification] (electronic lecture). Available at: www.ccas.ru/voron/download/MetricAlgs.pdf (accessed 24 November 2013). (in Russian)
3. Vorontsov K.V. Algoritmicheskie kompozicii [Algorithmic Compositions] (electronic lecture). Available at: http://www.ccas.ru/voron/download/Composition.pdf (accessed 24 November 2013). (in Russian)
4. D'yakonov A.G. (2013) Reshenie zadach analiza dannyh, osnovannoe na linejnoj kombinacii deformacij [Data Mining Problems Solving by Using of Linear Combinations of Deformations]. *Machine learning and data analysis*, vol. 1, no. 5, pp. 543–554. (in Russian)
5. Vorontsov K.V. (2010) Kombinatornaja teorija nadjozhnosti obuchenija po precedentam [The Combinatory Theory of Machine Learning Reliability] (PhD Thesis), Moscow: Dorodnicyn Computing Centre of RAS. (in Russian)
6. Rasin D. (2008) Neparametricheskaja ekonometrika: vvodnyj kurs [Non-parametric Econometrics: Introduction]. *Kvantil*, no. 4, pp. 7-56. (in Russian)
7. Parzen E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, no. 33, pp. 1065-1076.
8. Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, no. 27, pp. 832-837.