Информационные процессы, Том 16, № 1, 2016, стр. 72–89

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Формирование заказов торговой сети с помощью агрегирования специализированных алгоритмов прогнозирования[1]

В. В. Вьюгин, А.И. Шамсутдинов

Институт проблем передачи информации им. А.А.Харкевича, Российская академия наук, Москва, Россия

e-mail: vyugin@iitp.ru

Поступила в редколлегию 21.03.2016

Аннотация—В рамках математической теории машинного обучения рассматривается задача агрегации прогнозов специализированных экспертных стратегий. Под экспертными стратегиями понимаются алгоритмы, которые в режиме онлайн выдают последовательные прогнозы элементов временного ряда. Специализированные стратегии могут воздерживаться от прогнозов в некоторые моменты времени – они выдают прогнозы в соответствии с областью применимости той специфической модели предметной области, которая лежит в их основе. Предложен оптимальный алгоритм агрегирующий прогнозы таких экспертных стратегий в один прогноз. Оптимальность алгоритма заключается в том, что его суммарные потери в среднем асимптотически не больше чем потери любой стратегии прогнозирования на множестве моментов времени, когда она была активна. Получена верхняя оценка ошибки такого смешивания прогнозов – регрета агрегирующей стратегии. Оценки ошибки получены в наихудшем случае, когда не делается никаких предположений о механизме, лежащем в основе источника исходных данных. Проведено тестирование предложенного алгоритма на реальных данных о товарообороте торговой сети. Представлены численные результаты и оценки регрета.

КЛЮЧЕВЫЕ СЛОВА: онлайн обучение, агрегирующий алгоритм, эксперты специалисты, адаптивный регрет.

1. ВВЕДЕНИЕ

Теория предсказаний с использованием экспертных стратегий (Prediction with expert advice) является одним из разделов математической теории машинного обучения. В рамках этой теории строятся алгоритмы, объединяющие предсказания различных методов прогнозирования [1–5]. В качестве экспертных стратегий (или просто экспертов) могут выступать любые методы прогнозирования – это могут быть алгоритмы, результаты измерения природных процессов, экспертные оценки специалистов т.д. Каждая такая стратегия может использовать свою модель предметной области и может быть эффективной на определенных участках временного ряда.

В теории предсказаний с использование экспертных стратегий строятся методы агрегации специализированных прогнозирующих алгоритмов и даются теоретические гарантии их эффективности. Основной такой гарантией является ошибка обучения агрегирующего алгоритма – регрет. Характерная черта данного подхода в том, что оценки ошибки получаются в наихудшем случае, когда не делается никаких предположений о механизме источника исходных данных и способе работы специализированных прогнозирующих алгоритмов.

[1] Работа выполнена в ИППИ РАН за счет гранта Российского научного фонда, проект 14-50-00150.

Рассматривается задача прогнозирования будущего значения временного ряда некоторых показателей, которые генерируются последовательно в режиме онлайн источником, природа которого может быть неизвестна: на раундах (шагах) $t = 1, 2, \ldots$ некоторый источник генерирует исходы $y_1, y_2, \ldots$. В каждый момент времени t алгоритм предсказания, используя значения прошлых исходов $y_1, \ldots, y_{t-1}$ и различную доступную ему дополнительную информацию, выдает прогноз γ_t будущего исхода.

В данной работе используется набор известных методов предсказания значений временного ряда. В частности, это алгоритмы Ridge Regression, Lasso Regression, Bayesian Ridge Regression, Random Forest Regression и Gradient Boosting Regression. На каждом раунде игры каждый метод предсказания может выдавать свой прогноз показателя y_t. Каждый такой метод может использовать некоторую свою специальную модель источника. Поэтому на разных отрезках временного ряда различные методы прогнозирования могут иметь различную эффективность. Кроме этого, рассматривается постановка, при которой некоторые методы воздерживаются от прогнозов на некоторых шагах (или же эти прогнозы нежелательно использовать ввиду их неэффективности на прошлых раундах игры). Поэтому мы говорим о специализированных экспертах, такие эксперты в мировой литературе также называются спящими.

После того как исход y_t получен, качество прогноза каждого из алгоритмов оценивается величиной потерь. Потери на раунде t равны величине $\lambda(y_t, \gamma_t)$, где $\lambda(y, \gamma)$ – некоторая функция потерь. Потери суммируются, суммарные (кумулятивные) потери характеризуют эффективность работы предсказательного алгоритма. В практических приложениях раздела 4 будет использоваться выпуклая по γ функция потерь $\lambda(\gamma, y) = \lambda_1|\gamma - y|_+ + \lambda_2|y - \gamma|_+$, где λ_1 и λ_2 – положительные вещественные числа.

Мы рассматриваем задачу агрегации (смешивания) прогнозов специализированных экспертов в один единый прогноз. Наша цель – получить агрегированный алгоритм с наименьшими кумулятивными потерями. Более точно, кумулятивные потери агрегирующего алгоритма будут сравниваться с кумулятивными потерями наилучшего в прошлом алгоритма.

Согласно полученным теоретическим результатам, наша агрегирующая стратегия будет не сильно уступать любому из экспертов по качеству предсказаний.

Подробнее с теорией агрегирования предсказаний экспертов можно ознакомиться, например, в книгах [5, 6]. Практические применения представлены в работах [7, 8].

В работах [7, 9, 10] рассматриваются так называемые “спящие” экспертные стратегии или просто эксперты, которые могут воздерживаться от предсказания (быть неактивными или, как говорят, “спать”) на некоторых шагах. Мы называем также такие стратегии специализированными. Теоретические гарантии в этом случае распространяются только на те шаги, на которых эксперт не спит. Преимущество данного подхода в том, что новые эксперты в такой формулировке могут появляться в процессе работы алгоритма и предсказания некоторых экспертов могут отсутствовать в некоторые моменты времени t, что на практике случается достаточно часто (см. обсуждение в разделе 4).

В разделе 2 формулируется и изучается классический алгоритм Hedge с постоянным параметром обучения, предложенный в работе [4]. Проводится анализ ошибки обучения этого алгоритма – регрета. Алгоритм Hedge лежит в основе современных алгоритмов агрегации.

Основным параметром алгоритма Hedge является параметр обучения. Постоянный параметр обучения ведет к регрету, который растет линейно в зависимости от времени. Для получения более точных оценок на регрет используются переменные параметры обучения (соответствующие алгоритмы см. в [5] и [6]).

В разделе 3 обсуждается современный вариант алгоритма Hedge для переменного параметра обучения – алгоритм AdaHedge, впервые предложенный и изученный в работе [11].

Основным результатом данной работы является построение модификации этого алгоритма – алгоритм sAdaHedge для случая спящих экспертов. Соответствующий алгоритм и оценки его регрета представлены в разделе 3. Оценки регрета существенно используют результаты работы [11].

В разделе 4 приводятся результаты численных экспериментов проведенных с помощью алгоритма sAdaHedge на реальных исторических данных. Численные эксперименты и расчеты по прогнозированию временных рядов проводились на данных по ежедневному товарообороту некоторой сети–ритейлера.

Задача прогнозирования возникает из потребности отделов закупок, продаж и логистики иметь представление о предстоящих значениях продаж, чтобы увеличить маржинальность компании и минимизировать издержки. Аналогичные приложения для прогнозирования потребления электроэнергии были представлены в работе [7].

2. ЭКСПОНЕНЦИАЛЬНОЕ СМЕШИВАНИЕ ПОТЕРЬ

Рассмотрим формальную постановку задачи предсказания с использованием экспертных стратегий. Следующие понятия были введены в работе [12], алгоритм Hedge представлен в [4]. Более современная версия – алгоритм AdaHedge представлена в работе [11].

Имеется N – экспертных стратегий (экспертов). В алгоритме Hedge начальные веса экспертов определяются $p_{i,1} = \frac{1}{N}$.

На каждом шаге $t > 1$ алгоритм экспоненциального смешивания использует вектор весов экспертных стратегий $p_t = (p_{1,t}, \ldots, p_{N,t})$. Данные веса вычислены на предыдущем шаге $t-1$ в зависимости от эффективности предсказаний экспертов.

На шаге t алгоритм Hedge получает потери l_t^i всех экспертов $i = 1, \ldots, N$ и вычисляет свои потери $h_t = \sum\limits_{i=1}^{T} p_{i,t}^* l_t^i$, где

$$p_{i,t}^* = \frac{p_{i,t}}{\sum\limits_{j=1}^{N} p_{j,t}}$$

– нормированные веса экспертов. После этого подготавливаются веса экспертов $i = 1, \ldots, N$ для использования на следующем шаге

$$p_{i,t+1} = p_{i,t} e^{-\eta l_t^i},$$

где $\eta > 0$ – некоторый параметр алгоритма Hedge, называемый параметром обучения.

Кумулятивные потери алгоритма Hedge на шаге t определяются как $H_t = \sum\limits_{s=1}^{t} h_s$, кумулятивные потери эксперта i определяются нак $L_t^i = \sum\limits_{s=1}^{t} l_s^i$. Легко видеть, что $p_{i,t} = p_{i,1} e^{-\eta L_{t-1}^i}$.

Рассмотрим основные понятия, связанные с алгоритмом экспоненциального взвешивания с постоянным и переменным параметрами обучения.

Рассмотрим смешанные потери алгоритма (см. [11])

$$m_t = -\frac{1}{\eta} \ln \left(\sum_{s=1}^{N} p_{s,t}^* e^{-\eta l_t^s} \right)$$

Определим также кумулятивные смешанные потери $M_T = \sum\limits_{t=1}^{T} m_t$.

Определим $W_t = \sum\limits_{i=1}^{N} p_{i,t}$ для всех t. Из определения $p^*_{i,t} = \frac{p_{i,t}}{W_t}$.

Лемма 1. $M_T = -\frac{1}{\eta} \ln W_{T+1} = -\frac{1}{\eta} \ln \sum\limits_{i=1}^{N} p_{i,1} e^{-\eta L^i_T}$.

Доказательство. Действительно,

$$M_T = \sum_{t=1}^{T} m_t = \sum_{t=1}^{T} -\frac{1}{\eta} \ln \frac{1}{W_t} \sum_{i=1}^{N} p_{i,t} e^{-\eta l^i_t} = \sum_{t=1}^{T} -\frac{1}{\eta} \ln \frac{1}{W_t} \sum_{i=1}^{N} p_{i,t+1} =$$
$$-\frac{1}{\eta} \sum_{t=1}^{T} \ln \frac{W_{t+1}}{W_t} = -\frac{1}{\eta} \ln \prod_{t=1}^{T} \frac{W_{t+1}}{W_t} = -\frac{1}{\eta} \ln W_{T+1} =$$
$$-\frac{1}{\eta} \ln \sum_{i=1}^{N} p_{i,T+1} = -\frac{1}{\eta} \ln \sum_{i=1}^{N} p_{i,1} e^{-\eta L^i_T}.$$

□

Так как сумма не меньше чем любое слагаемое, из второго равенства леммы 1 получаем следующие неравенства:

Лемма 2. *Кумулятивные смешанные потери связаны с кумулятивными смешанными потерями произвольного эксперта i следующим образом*

$$M_T \le -\frac{1}{\eta} \ln p_{i,T+1} = L^i_T + \frac{1}{\eta} \ln \frac{1}{p_{i,1}} = L^i_T + \frac{\ln N}{\eta}. \tag{1}$$

Следующие вспомогательные леммы являются прямыми аналогами соответствующих лемм из работы [11]. Мы приводим их формулировки и доказательства для того, чтобы отследить специфику их применения для случая экспертов специалистов.

Введем обозначения m^η_t и M^η_T для соответствующих величин m_t и M_T определенных при значении параметра обучения равном η. Пусть также $l^-_t = \min_{1 \le i \le N} l^i_t$ и $l^+_t = \max_{1 \le i \le N} l^i_t$.

Лемма 3. *Для любого постоянного параметра обучения $\eta \in (0, \infty]$*

1. $l^-_t \le m^\eta_t \le \sum\limits_{i=1}^{N} p^*_{i,t} l^i_t \le l^+_t$.
2. *Смешанные потери M^η_t есть невозрастающая функция по η, т.е $M^\eta_T \le M^\mu_T$ при $\mu \ge \eta$.*

Доказательство. Неравенства $l^-_t \le m^\eta_t$ и $\le \sum\limits_{i=1}^{N} p^*_{i,t} l^i_t \le l^+_t$ очевидны.

Докажем неравенство $m^\eta_t \le \sum\limits_{i=1}^{N} p^*_{i,t} l^i_t$. Для этого воспользуемся неравенством

$$\varphi(EX) \le E\varphi(X), \tag{2}$$

где X – случайная величина, E – символ математического ожидания, φ – выпуклая функция.

Пусть X есть случайная величина принимающая значения l_t^j с вероятностями $p_{t,j}^*$. Пусть $p_t^* = \{p_{j,t}^* : j = 1, \ldots, N\}$, пусть также E – символ математического ожидания. Тогда

$$\begin{aligned}
e^{-\eta E_{j\sim p_t^*}[l_t^j]} &\le E_{j\sim p_t^*}[e^{-\eta l_t^j}];\\
-\eta E_{j\sim p_t^*}[l_t^j] \le \ln E_{j\sim p_t^*}[e^{-\eta l_t^j}];& -\ln E_{j\sim p_t^*}[e^{-\eta l_t^j}] \le \eta E_{j\sim p_t^*}[l_t^j];\\
-\frac{1}{\eta}\ln E_{j\sim p_t^*}[e^{-\eta l_t^j}] &\le E_{j\sim p_t^*}[l_t^j];\\
-\frac{1}{\eta}\ln\left(\sum_{j=1}^N p_{j,t}^* e^{-\eta l_t^j}\right) &\le \sum_{j=1}^N p_{j,t}^* l_t^j;\\
m_t^\eta &\le \sum_{i=1}^N p_{i,t}^* l_t^i.
\end{aligned}$$

Рассмотрим $\eta \le \mu$ и воспользуемся неравенством (2). По лемме 1 и выпуклости экспоненты имеем

$$\begin{aligned}
M_T^\eta = -\frac{1}{\eta}\ln\sum_{i=1}^N p_{i,1}e^{-\eta L_T^i} = -\frac{1}{\eta}\ln\sum_{i=1}^N p_{i,1}\left(e^{-\mu L_T^i}\right)^{\frac{\eta}{\mu}} \ge\\
-\frac{1}{\eta}\ln\left(\sum_{i=1}^N p_{i,1}e^{-\mu L_T^i}\right)^{\frac{\eta}{\mu}} = -\frac{1}{\mu}\ln\sum_{i=1}^N p_{i,1}e^{-\mu L_T^i} = M_T^\mu.
\end{aligned}$$

Таким образом, смешанные потери есть невозрастающая функция по η. □

Согласно работе [11], в алгоритме AdaHedge используются переменный параметр обучения η_t, а нормированные веса экспертов заново определяются на каждом шаге как $p_{i,1} = 1/N$ и при $t > 1$

$$p_{i,t} = \frac{e^{-\eta_t L_{t-1}^i}}{\sum\limits_{j=1}^N e^{-\eta_t L_{t-1}^j}}$$

для всех i, где L_{t-1}^i – кумулятивные потери эксперта i за предыдущие $t-1$ шагов. Смешанные потери на шаге t в этом случае равны

$$m_t = -\frac{1}{\eta_t}\ln\left(\sum_{s=1}^N p_{i,t}e^{-\eta_t l_{i,t}}\right) = -\frac{1}{\eta_t}\ln\left(\sum_{s=1}^N \frac{e^{-\eta_t L_t^i}}{\sum\limits_{j=1}^N e^{-\eta_t L_{t-1}^j}}\right). \tag{3}$$

Из определения

$$m_t = M_t^{\eta_t} - M_{t-1}^{\eta_t}. \tag{4}$$

Лемма 4. *Пусть $M_T^{\eta_T}$ – кумулятивные смешанные потери с постоянным параметром обучения $\eta = \eta_T$. Тогда кумулятивные смешанные потери M_T алгоритма AdaHedge с переменным параметром обучения η_t (где $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_T$) на каждом шаге t удовлетворяет неравенству $M_T \le M_T^{\eta_T}$ для всех T.*

Доказательство. Так как по лемме 3 величина M^η есть невозрастающая функция от η, а также выполнено (4), будет

$$M_T = \sum_{t=1}^{T} m_t = \sum_{t=1}^{T} M_t^{\eta_t} - M_{t-1}^{\eta_t} \le \sum_{t=1}^{T} M_t^{\eta_t} - M_{t-1}^{\eta_{t-1}} = M_T^{\eta_T},$$

что и требовалось доказать. □

Отсюда по леммам 1 и 2 получаем

$$M_T \le M_T^{\eta_T} = \sum_{t=1}^{T} m_t^{\eta_T} = -\frac{1}{\eta_T} \ln W_{T+1}^{\eta_T} = -\frac{1}{\eta_T} \ln \left(\sum_{i=1}^{N} p_{i,1} e^{-\eta_T L_T^i} \right) \le L_T^i + \frac{\ln N}{\eta_T} \tag{5}$$

для любого i.

3. АЛГОРИТМ sAdaHedge

Как и в предыдущем пункте, будем рассматривать ту же постановку задачи, но с той разницей, что на шаге t эксперты разобьются на два множества – активных и спящих. Множество активных экспертов в момент времени t примем за $E_t \subset \{1, \ldots, N\}$. На шаге t потери эксперта i равны l_t^i при $i \in E_t$ и не определены при $i \notin E_t$. Кумулятивные потери эксперта специалиста $\sum\limits_{t:i\in E_t}^{T} l_t^i$.

Будем использовать адаптивный параметр обучения η_t, который будет определяться в процессе выполнения алгоритма.

Модифицируем алгоритм AdaHedge для случая экспертов специалистов. Используем идею из работы [10] – искусственным образом приписываем каждому спящему на шаге t эксперту $i \notin E_t$ потери равные смешанным потерям m_t. На момент приписывания эти потери являются виртуальными, так как сама величина m_t должна вычисляться через потери всех экспертов. При этом мы получаем уравнение, содержащее переменную m_t, из которого мы и получаем новые формулы для вычисления новых весов уже только активных экспертов $p_{i,t}^E$. Величина m_t вычисляется через эти веса и потери активных на шаге t экспертов.

На шаге t получаем список E_t активных экспертов специалистов. Получаем потери l_t^i каждого эксперта специалиста $i \in E_t$. Для спящего на шаге t эксперта $i \notin E_t$ предполагаем, что $l_t^i = m_t$. Тогда из определения смешанных потерь

$$m_t = -\frac{1}{\eta} \ln \sum_{i=1}^{N} p_{i,t} e^{-\eta_t l_t^i}$$

получаем уравнение относительно m_t

$$e^{-\eta_t m_t} = \sum_{i=1}^{N} p_{i,t} e^{-\eta_t l_t^i} = \sum_{i\in E_t} p_{i,t} e^{-\eta_t l_t^i} + \sum_{i\notin E_t} p_{i,t} e^{-\eta_t m_t} = \sum_{i\in E_t} p_{i,t} e^{-\eta_t l_t^i} + e^{-\eta_t m_t} \left(1 - \sum_{i\in E_t} p_{i,t} \right).$$

Отсюда выражаем m_t и находим новые веса экспертов специалистов и новое представление смешанных потерь m_t через эти веса

$$m_t = -\frac{1}{\eta_t} \ln \frac{\sum\limits_{i \in E_t} p_{i,t} e^{-\eta_t l_t^i}}{\sum\limits_{j \in E_t} p_{j,t}} = \sum_{i \in E_t} p_{i,t}^{E_t} e^{-\eta_t l_t^i}, \tag{6}$$

где

$$p_{i,t}^{E_t} = \frac{p_{i,t}}{\sum\limits_{j \in E_t} p_{j,t}}.$$

Результирующий алгоритм представлен ниже.

Алгоритм sAdaHedge

Входные параметры: начальные веса экспертов $p_{i,1} = 1/N$, $L_0^i = 0$,
$i = 1, \ldots, N$, $\Delta_0 = 0$, $H_0 = 0$

FOR $t = 1, 2, \ldots$

Полагаем параметр обучения равным $\eta_t = \frac{\ln N}{\Delta_{t-1}}$

Получаем множество активных (неспящих) экспертов специалистов E_t

Определим $p_{i,t}^{E_t} = \dfrac{p_{i,t}}{\sum\limits_{i \in E_t} p_{i,t}}$ при $i \in E_t$

Получаем потери экспертов специалистов l_t^i при $i \in E_t$

Вычисляем потери алгоритма sAdaHedge $h_t = \sum\limits_{i \in E_t} p_{i,t}^{E_t} l_t^i$

Вычисляем кумулятивные потери всех экспертов $1 \le i \le N$ и алгоритма $L_t^i = L_{t-1}^i + l_t^i$, где l_t^i получены при $i \in E_t$ и $l_t^i = m_t$ при $i \notin E_t$. Здесь

$$m_t = -\frac{1}{\eta_t} \ln \frac{\sum\limits_{i \in E_t} p_{i,t} e^{-\eta_t l_t^i}}{\sum\limits_{j \in E_t} p_{j,t}}$$

$H_t = H_{t-1} + h_t$

$\Delta_t = \Delta_{t-1} + \delta_t$, где $\delta_t = h_t - m_t$

Вычисляем веса всех экспертов $1 \le i \le N$ для шага $t + 1$

$$p_{i,t+1} = \frac{e^{-\eta_t L_{i,t}}}{\sum\limits_{j=1}^{N} e^{-\eta_t L_{j,t}}},$$

ENDFOR

Для оценки качества предсказаний введем понятие регрета, который сравнивает кумулятивную ошибку алгоритма и эксперта с минимальными потерями.

Определим расширенные кумулятивные потери эксперта i с учетом виртуальных потерь $L_T^i = \sum\limits_{t=1}^{T} l_t^i$.

Регрет алгоритма AdaHedge (относительно эксперта i) определяется как $R_T^i = H_T - L_T^i$, где $H_T = \sum\limits_{t=1}^{T} h_t$ – кумулятивные потери алгоритма за T шагов, $L_T^i = \sum\limits_{t=1}^{T} l_t^i$ – расширенные кумулятивные потери эксперта i за T шагов. Из определения для произвольного i выполнено

$$R_T^i = \sum_{t=1}^{T} h_t - L_T^i = \sum_{t=1}^{T}(m_t + \delta_t) - L_T^i = M_T - L_T^i + \Delta_T$$

Чтобы найти точную оценку на регрет, найдем отдельно оценки для $\max_{1\le i\le N}(M_T - L_T^i)$ и Δ_T.

По лемме 3 $\delta_t = h_t - m_t \ge 0$, поэтому $\Delta_{t-1} \le \Delta_t$ для всех t. Согласно оценке (5) и определению параметра обучения

$$M_T \le L_T^i + \frac{\ln N}{\eta_T} = L_T^i + \Delta_{T-1} \le L_T^i + \Delta_T.$$

Отсюда получаем оценку для регрета (относительно эксперта i) алгоритма AdaHedge

$$R_T^i \le L_T^i + \frac{\ln N}{\eta} - L_T^i + \Delta_T = 2\Delta_T.$$

Регрет алгоритма sAdaHedge (относительно эксперта i) определяется как

$$\tilde{R}_T^i = \sum_{t:i\in E_t} h_t - \sum_{t:i\in E_t} l_t^i,$$

где суммирование потерь производится только по тем шагам на которых эксперт i был активен.

Выразим оценку регрета $\tilde{R}_T^i$ алгоритма sAdaHedge относительно произвольного эксперта специалиста i на тех шагах, на которых он был активен. Этот регрет не превосходит регрет R_T^i алгоритма AdaHedge, который вычислен с учетом виртуальных потерь экспертов в те моменты времени, когда они спали.

Лемма 5. *Для произвольного i и всех T*

$$\tilde{R}_T^i \le R_T^i \le 2\Delta_T.$$

Доказательство. Действительно, для произвольного i будет

$$\begin{gathered}\tilde{R}_T^i = \sum_{t:i\in E_t} h_t - \sum_{t:i\in E_t} l_t^i = \\ \sum_{t:i\in E_t} h_t + \sum_{t:i\notin E_t} m_t - \sum_{t:i\in E_t} l_t^i - \sum_{t:i\notin E_t} m_t \le \\ \sum_{t:i\in E_t} h_t + \sum_{t:i\notin E_t} h_t - \sum_{t:i\in E_t} l_t^i - \sum_{t:i\notin E_t} l_t^i = \\ \sum_{t=1}^{T} h_t - \sum_{t=1}^{T} l_t^i = R_T^i \le 2\Delta_T.\end{gathered}$$

Здесь мы использовали представление потерь h_t алгоритма и представление (6) смешанных потерь m_t через веса спящих экспертов, а также неравенство $m_t \le h_t$ для всех t, которое доказано в лемме 3. □

Осталось найти оценку на Δ_T.

Напомним, что $l_t^- = \min_{i\in E_t} l_t^i$ и $l_t^+ = \max_{i\in E_t} l_t^i$. Так как потери спящих экспертов есть смешанная комбинация (6) потерь активных экспертов, выполнено $l_t^- = \min_{1\le i\le N} l_t^i$ и $l_t^+ = \max_{1\le i\le N} l_t^i$.

Определим $s_t = l_t^+ - l_t^-$ и $S_T = \max_{1\le t\le T}(l_t^+ - l_t^-)$.

В работе [11] были получены оценки регрета алгоритма AdaHedge без учета спящих экспертов. Мы обобщим первую часть этой оценки на случай экспертов специалистов.

В алгоритме sAdaHedge мы приписываем спящим экспертам потери агрегирующего алгоритма, которые вычисляются по специальной формуле через потери активных на шаге t экспертов. Заметим, что смешанные потери m_t имеют два представления через старые и новые веса (6). Поэтому формально основная часть анализа алгоритма sAdaHedge, связанная со смешанными потерями, совпадает с анализом алгоритма AdaHedge из работы [11]. Тем не менее, имеются некоторые важные отличия. По этой причине и для полноты изложения приводим модифицированные оценки следуя схеме работы [11].

Приведем необходимые оценки с помощью неравенства Бернштейна через кумулятивную дисперсию потерь $V_T = \sum\limits_{t=1}^{T} v_t$, где

$$v_t = Var_{j\sim p_t^{E_t}}[l_t^j] = E_{j\sim p_t^{E_t}}[(l_t^j - E_{j\sim p_t^{E_t}}[(l_t^j)])^2] = \sum_{j\in E_t} p_t^{E_t}(l_t^j - h_t)^2.$$

Обозначим $s_t = l_t^+ - l_t^-$.

Лемма 6. *Смешанная разность δ_t удовлетворяет:*

$$\delta_t \le \frac{e^{s_t\eta_t} - 1 - s_t\eta_t}{\eta_t s_t^2} v_t. \tag{7}$$

Доказательство. Докажем это выражение, используя неравенство Бернштейна (см. леммы A3–A5 из [5]).

Предварительно сформулируем это неравенство для произвольной случайной величины $X \in (-\infty, 1)$, где $EX = 0$, $EX^2 = \sigma^2$. Тогда для произвольного $\eta > 0$ будет

$$\ln Ee^{\eta X} \le \sigma^2(e^\eta - \eta - 1).$$

В дальнейшем также будет использоваться следующее неравенство. Пусть $X \in [0, 1]$ – случайная величина и $\sigma = \sqrt{EX^2 - (EX)^2}$. Тогда для произвольного $\eta > 0$ будет

$$\ln E[e^{-\eta(X-EX)}] \le \sigma^2(e^\eta - \eta - 1).$$

Пусть $p_t^{E_t} = (p_{1,t}^{E_t}, \ldots, p_{N,t}^{E_t})$ – распределение экспертов активных на шаге t. В качестве X рассмотрим случайную величину, которая принимает значения l_t^j с вероятностями $p_{j,t}^{E_t}$, где $j = 1, \ldots, N$. Преобразуем ее так, чтобы она лежала на отрезке $[0, 1]$: $X_t^j = \dfrac{l_t^j - l_t^-}{s_t}$. Тогда неравенство Бернштейна можно записать следующим образом:

$$\ln E_{j\sim p_t^{E_t}}\left(e^{-\eta\left(X_t^j - EX_t^j\right)}\right) \le \sigma^2\left(e^\eta - 1 - \eta\right)$$

для всех $\eta > 0$. Перепишем это неравенство более детально при $\eta = s_t\eta_t$:

$$\ln\left(\sum_{j\in E_t} p_{t,j}^{E_t} e^{-s_t\eta_t\left(\frac{l_t^j - l_t^-}{s_t} - \sum\limits_{j\in E_t} p_{i,t}^{E_t}\frac{l_t^i - l_t^-}{s_t}\right)}\right) = \ln\left(\sum_{j\in E_t} p_{t,j}^{E_t} e^{-\eta_t\left(l_t^j - l_t^- - \sum\limits_{j\in E_t} p_{i,t}^{E_t}\left(l_t^i - l_t^-\right)\right)}\right) =$$

$$= \ln\left(\frac{\sum\limits_{j\in E_t} p_{j,t}^{E_t} e^{-\eta_t\left(l_t^j - l_t^-\right)}}{e^{-\eta_t\sum\limits_{j\in E_t} p_{i,t}^{E_t}\left(l_t^j - l_t^-\right)}}\right) = \ln\left(\frac{\sum\limits_{j\in E_t} p_{j,t}^{E_t} e^{-\eta_t l_t^j}}{e^{-\eta_t\sum\limits_{j\in E_t}^{N} p_{j,t}^{E_t} l_t^i}}\right) =$$

$$\ln\sum_{j\in E_t} p_{j,t}^{E_t} e^{-\eta_t l_t^j} + \eta_t\sum_{j\in E_t} p_{j,t}^{E_t} l_t^j = \eta_t(h_t - m_t) = \eta_t\delta_t \le \sigma^2\left(e^{s_t\eta_t} - 1 - s_t\eta_t\right) =$$

$$Var_{j\sim p_t^{E_t}}[X_t^j]\left(e^{s_t\eta_t} - 1 - s_t\eta_t\right) = \frac{1}{s_t^2} Var_{j\sim p_t^{E_t}}[l_t^j]\left(e^{s_t\eta_t} - 1 - s_t\eta_t\right).$$

Откуда получаем наше исходное неравенство:

$$\delta_t \le \frac{e^{s_t\eta_t} - 1 - s_t\eta_t}{\eta_t s_t^2} v_t.$$

□

Запишем (7) также в виде

$$\delta_t \le \frac{g(s_t,\eta_t}{s_t} v_t, \text{ где } g(x) = \frac{e^x - x - 1}{x}. \tag{8}$$

Воспроизведем оценку на кумулятивную смешанную разность Δ_T из работы [11] с помощью неравенства Бернштейна.

Лемма 7. *Кумулятивная смешанная разность Δ_T удовлетворяет:*

$$(\Delta_T)^2 \le V_T \ln N + \left(\frac{2}{3}\ln N + 1\right) S_T\Delta_T,$$

где $S_T = \max_{1\le t\le T}(l_t^+ - l_t^-)$.

Доказательство. Распишем квадрат кумулятивной смешанной разности:

$$(\Delta_T)^2 = \sum_{t=1}^{T}\left(\Delta_t^2 - \Delta_{t-1}^2\right) = \sum_{t=1}^{T}\left(\left(\Delta_{t-1} + \delta_t\right)^2 - \Delta_{t-1}^2\right) = \sum_{t=1}^{T}\left(2\delta_t\Delta_{t-1} + \delta_t^2\right) =$$

$$\sum_{t=1}^{T}\left(2\delta_t\frac{\ln N}{\eta_t} + \delta_t^2\right) \le \sum_{t=1}^{T}\left(2\delta_t\frac{\ln N}{\eta_t} + s_t\delta_t\right) \le 2\ln N\sum_{t=1}^{T}\frac{\delta_t}{\eta_t} + S_T\Delta_T l. \tag{9}$$

Получим оценку $\dfrac{\delta_t}{\eta_t}$ с помощью неравенства (8)

$$v_t \ge \frac{\delta_t s_t}{2g\left(s_t\eta_t\right)} = \frac{\delta_t}{\eta_t} + A,$$

$$A = \frac{\eta_t s_t\delta_t - 2g\left(s_t\eta_t\right)\delta_t}{2g\left(s_t\eta_t\right)\eta_t} = \frac{s_t\left(\eta_t^2\delta_t s_t^2 - 2\delta_t\left(e^{s_t\eta_t} - s_t\eta_t - 1\right)\right)}{2s_t\eta_t\left(e^{s_t\eta_t} - s_t\eta_t - 1\right)} =$$

$$= s_t\delta_t \frac{\frac{1}{2}(s_t\eta_t)^2 - e^{s_t\eta_t} - s_t\eta_t - 1}{s_t\eta_t\left(e^{s_t\eta_t} - s_t\eta_t - 1\right)} = \varphi(s_t\eta_t)\, s_t\delta_t,$$

где $\varphi = \frac{e^x - \frac{1}{2}x^2 - x - 1}{xe^x - x^2 - x}$.

Найдем оценку на $\varphi(x)$, разложив ее в ряд Тейлора:

$$\frac{e^x - \frac{1}{2}x^2 - x - 1}{xe^x - x^2 - x} \sim \frac{1 + x + \frac{x^2}{2} - \frac{x^2}{2} - x - 1 + \frac{x^3}{6} + o(x)}{\frac{x^3}{2} + o(x)} \sim \frac{\frac{x^3}{6} + o(x)}{\frac{x^3}{2} + o(x)} \sim \frac{1}{3}.$$

Откуда

$$\frac{\delta_t}{\eta_t} \le \frac{1}{3}s_t\delta_t + \frac{1}{2}v_t. \tag{10}$$

Напомним, что $V_T = \sum_{t=1}^{T} v_t$.

Подставив оценку (10) в неравенство (9) и просуммировав, получим:

$$(\Delta_T)^2 \le V_T \ln N + \left(\frac{2}{3}\ln N + 1\right) S_T\Delta_T.$$

□

Напомним, что $R_T^i = H_T^i - L_T^i$, где $H_T = \sum_{t=1}^{T} h_t$ – кумулятивные потери потери алгоритма AdaHedge за T шагов, $L_T^i = \sum_{t=1}^{T} l_t^i$ – расширенные кумулятивные потери потери эксперта i за T шагов (с учетом виртуальных потерь). Соответствующий регрет с учетом спящих экспертов: $\tilde{R}_T^i = \sum_{t:i\in E_t} h_t - \sum_{t:i\in E_t} l_t^i$. По лемме 5 $\tilde{R}_T^i \le R_T^i$ для всех i.

Оценим сверху регрет R_T^i, а тем самым и $\tilde{R}_T^i$, используя лемму 7.

Теорема 1. *Для произвольного i*

$$\tilde{R}_T^i \le 2\sqrt{V_T \ln N} + S_T\left(\frac{4}{3}\ln N + 2\right), \tag{11}$$

где $V_T = \sum_{t=1}^{T} Var_{j\sim p_t^{E_t}}[l_t^j]$.

Доказательство. Воспользуемся леммой 6:

$$(\Delta_T)^2 \le a + b\Delta_T,$$

где $a = V_T \ln N \ge 0$, $b = S_T\left(\frac{2}{3}\ln N + 1\right) \ge 0$.

Разрешим это неравенство относительно Δ_T:

$$\Delta_T \le \frac{1}{2}b + \frac{1}{2}\sqrt{b^2 + 4a} \le \frac{1}{2}b + \frac{1}{2}\left(\sqrt{b^2} + \sqrt{4a}\right) = \sqrt{a} + b.$$

Подставив это в утверждение леммы 7, получим:

$$R_T^i \le 2\sqrt{a} + 2b = 2\sqrt{V_T \ln N} + S_T\left(\frac{4}{3}\ln N + 2\right).$$

□

До настоящего времени мы рассматривали задачу прогнозирования в самом общем виде, как задачу построения оптимального смешивания абстрактных потерь экспертных стратегий. Данная постановка легко применяется к случаю получения агрегированных предсказаний на основе предсказаний экспертов.

Эксперты $i \in E_t$ в режиме онлайн на каждом шаге t делают предсказания $f_{i,t}$.[1] Алгоритм sAdaHedge вычисляет свое предсказание

$$\gamma_t = \sum_{i \in E_t} p_{i,t}^{E_t} f_{i,t},$$

где $p_{i,t}^{E_t}$ – модифицированные веса активных экспертов из протокола алгоритма sAdaHedge. После этого некоторый источник выдает исход y_t. Эксперты вычисляют свои потери используя некоторую функцию потерь $\lambda(y, \gamma)$, а именно, $l_t^i = \lambda(y_t, f_{i,t})$ при $i \in E_t$. Потери алгоритма равны $\lambda(y_t, \gamma_t)$. Мы предполагаем выпуклость функции $\lambda(y, \gamma)$ по γ. Отсюда $\lambda(y_t, \gamma_t) \le \sum\limits_{i \in E_t} p_{i,t}^{E_t} l_t^i = h_t$. Для кумулятивных потерь за первые T шагов имеем

$$\sum_{t:i \in E_t}^{T} \lambda(y_t, \gamma_t) \le \sum_{t:i \in E_t} h_t.$$

Отсюда следует, что оценка (11) теоремы 1 выполнена и для регрета вычисленного на основе потерь от прогнозов экспертов.

4. РЕЗУЛЬТАТЫ ЧИСЛЕННЫХ ЭКСПЕРИМЕНТОВ

Численные эксперименты и расчеты по прогнозированию временных рядов проводились на данных по ежедневному товарообороту немецкой сети–ритейлера. Задача возникает из потребности отделов закупок, продаж и логистики иметь представление о предстоящих значениях продаж, чтобы увеличить маржинальность компании и минимизировать издержки. Для удобства пользования данные были нормированы на отрезок $[0, 1]$.

Под экспертными стратегиями брались алгоритмы, которые в режиме онлайн выдавали последовательные прогнозы элементов временного ряда. Каждый из алгоритмов был проанализирован на предмет «хорошего» прогноза. Вследствие чего в некоторые моменты времени t его не включали в множество активных экспертов E_t.

В качестве экспертов специалистов были рассмотрены наиболее популярные и применяемые на практике алгоритмы. Это Ridge Regression, Lasso Regression, Bayesian Ridge Regression, Random Forest Regression и Gradient Boosting Regression. Подробнее с этими алгоритмами можно ознакомиться в книге [13]. Для контроля переобучения была проведена кросс-валидация всех алгоритмов. Рассмотрим каждый из этих алгоритмов в отдельности.

Ridge Regression есть модификация линейной регрессии на случай, когда переменные коррелируют друг с другом (т.е. имеет место мультиколлинеарность).

Lasso Regression. Метод заключается во введении ограничения на норму вектора весов модели, что приводит к обращению в нуль некоторых коэффициентов. Повышается устойчивость задачи в случае большого числа обусловленности матрицы признаков. Метод позволяет «отбирать» признаки, оказывающие наибольшее влияние на вектор ответов.

Bayesian Ridge Regression похож на Ridge Regression, но с тем отличием, что вектор весов имеет априорное нормальное распределение, где параметры подчиняются Гамма–распределению.

[1] Эксперты специалисты могут воздерживаться от предсказаний.

Random Forest Regression. Идея метода основана на бэггинге и построении случайных подпространств деревьев. Все деревья строятся независимо друг от друга. Итоговое значение для регрессии получается путем усреднения каждого дерева комитета.

Gradient Boosting Regression есть обобщение алгоритма бустинга, в основе алгоритма лежит последовательное уточнение функции, представляющей собой линейную комбинацию базовых моделей, с тем чтобы минимизировать функцию потерь.

Вся выборка данных поделена на обучающую и тестовую в соотношении 80% и 20%, соответственно. Обучающая выборка, в свою очередь, поделена на обучение самих экспертов специалистов и обучение алгоритма sAdaHedge. После анализа временного ряда и каждого из экспертов специалистов было установлено, что некоторые алгоритмы хорошо работают на понижение/повышение прогнозов или в определенные дни недели. Это явление можно наблюдать на Рис. 1.

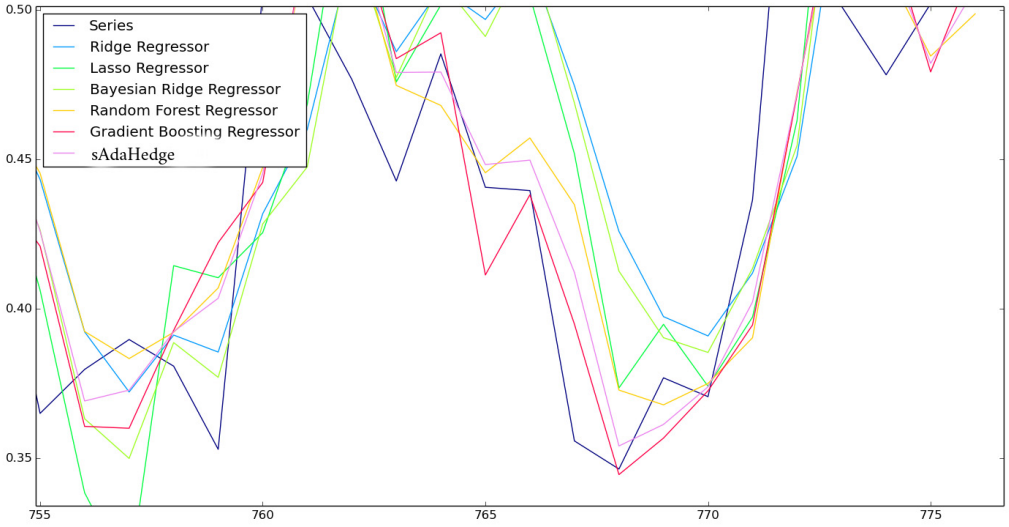

Рис. 1. График ряда и прогнозов экспертов.

Вследствие чего алгоритм множество активных экспертов E_t составлялось таким образом, чтобы математическое ожидание ошибки на обучающей выборке было наименьшим. Проведено усреднение потерь экспертов специалистов по дням недели. Анализ построения множества E_t для контрольной выборки проходил на этапе обучения алгоритмов экспертов и на этапе обучения алгоритма sAdaHedge. Такой подход построения множества активных экспертов E_t связан с тем, что часть экспертов специалистов лучше приспосабливается к выбросам, т.е. к тем значениям, которые не характерны для данного ряда. Другая же часть имеет маленькую ошибку в моменты стабильного поведения ряда.

В качестве функции потерь (меры ошибки) была рассмотрена несимметричная функция

$$\lambda(\gamma, y) = \lambda_1|\gamma - y|_+ + \lambda_2|y - \gamma|_+,$$

где γ — прогнозы экспертов специалистов, y — истинное значение ряда, λ_1, λ_2 — коэффициенты, определяемые экспертным мнением ритейлера–сети.

Такой вид функции потерь был выбран не случайно. Абсолютное значение между прогнозом и истинным значением отражает потери магазина в денежном эквиваленте, что имеет прямой

экономический смысл. Различные коэффициенты от недопрогноза и перепрогноза показывают убытки стоимости хранения товаров на складе, порчи товара по истечению срока годности и от дефицита, что несет в себе различный уровень потерь, важный для ведения бизнеса.

Рассмотрим на Рис. 2 кумулятивные потери самих экспертов специалистов за все время прогноза и потери алгоритма sAdaHedge.

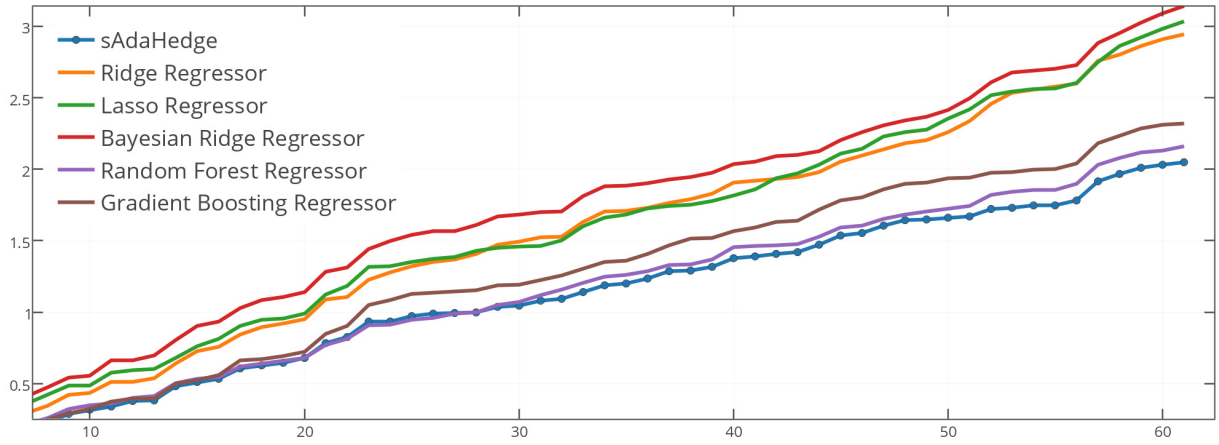

Рис. 2. Кумулятивные потери всех алгоритмов, в том числе и sAdaHedge, на всем интервале прогноза.

Видно, что, начиная с некоторого момента, кумулятивные потери алгоритма sAdaHedge становятся меньше лидера за счет правильного и эффективного построения множества активных экспертов E_t в каждый момент времени t, когда эксперт был активным.

Регрет агрегирующего алгоритма, начиная с некоторого момента времени t, становится меньше нуля, тем самым показывая, что алгоритм sAdaHedge не уступает, а в данном случае даже превосходит лучшего из экспертов специалистов по величине ошибки. Изменение регрета алгоритма sAdaHedge показан на Рис. 3.

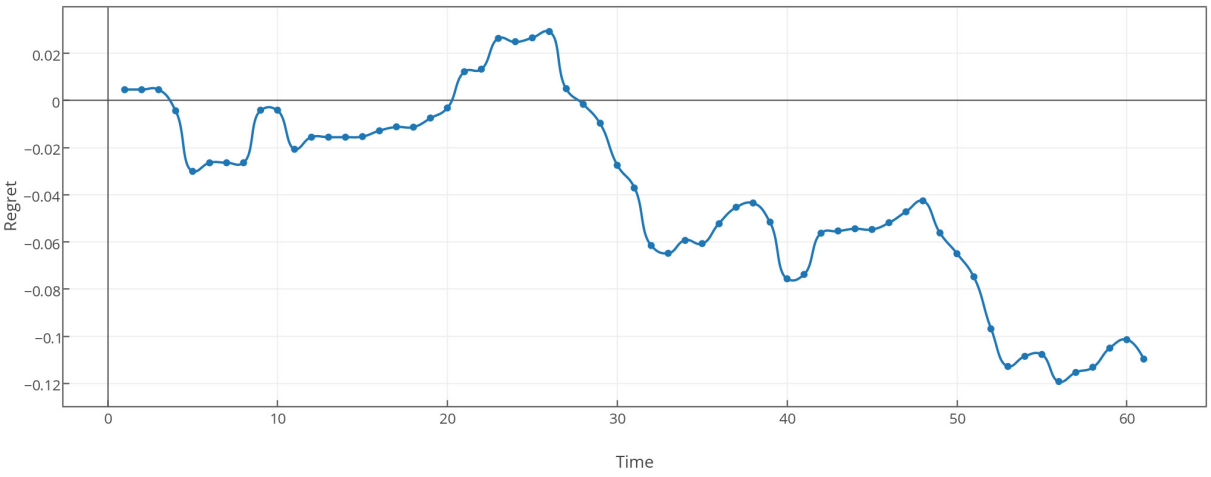

Рис. 3. Регрет алгоритма sAdaHedge.

В работе рассматривается еженедельный товарооборот розничной сети, поэтому можно заключить, что присутствует некая сезонность данных. В качестве основы таких отрезков были рассмотрены недели. Поэтому множество активных экспертов E_t на каждой итерации строилось таким образом, что в него не попадали те из экспертов, кто на стадии обучения в этот момент времени имел большую долю ошибки. Активность экспертов специалистов на тестовой выборке представлена на Рис. 4.

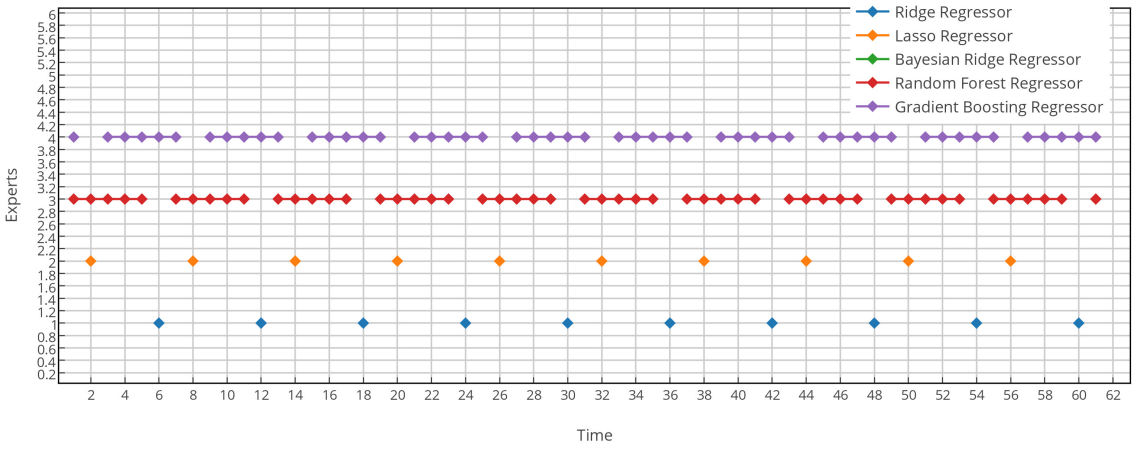

Рис. 4. Активность экспертов специалистов в каждый момент времени.

Сравнивая ошибки (Рис. 5–8) алгоритма sAdaHedge с каждым алгоритмом из множества экспертов специалистов в моменты времени, когда он был активен, можно заметить, что потери алгоритма sAdaHedge в большинстве случаев не превосходят потерь каждого из алгоритмов. На каждой итерации потери агрегирующего алгоритма близки к потерям эксперта специалиста в случае меньших потерь у последних, т.е. как и было показано в теоретических оценках — суммарные потери в среднем асимптотически не больше, чем потери любой стратегии прогнозирования на множестве моментов времени, когда она была активна. Итоговые кумулятивные потери всех алгоритмов отображены в Таблице 1.

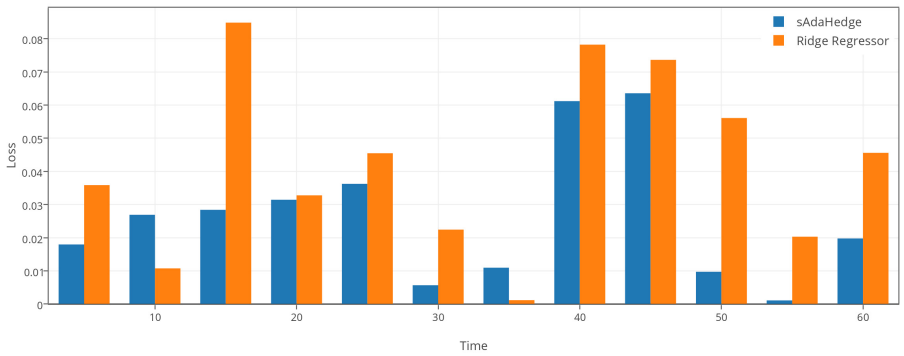

Рис. 5. Сравнение потерь алгоритма sAdaHedge и Ridge Regression в моменты времени t, когда он был активен.

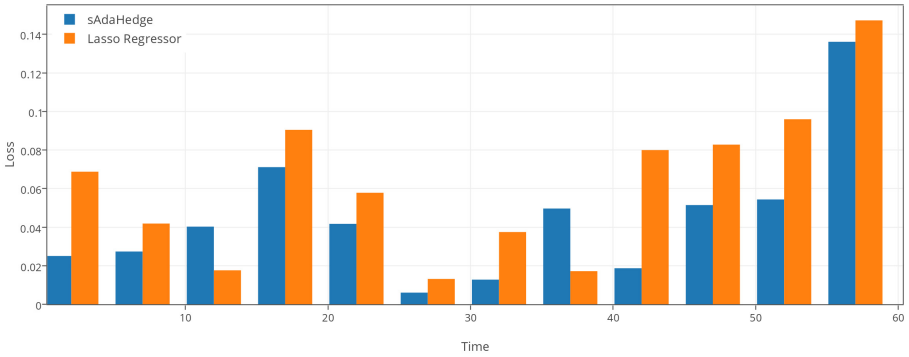

Рис. 6. Сравнение потерь алгоритма sAdaHedge и Lasso Regression в моменты времени t, когда он был активен.

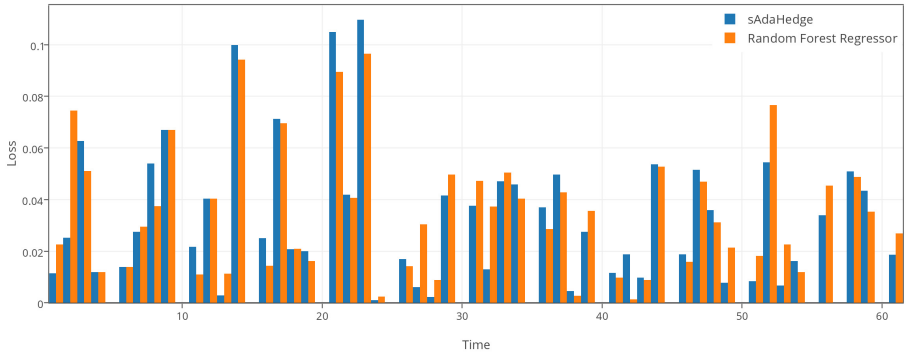

Рис. 7. Сравнение потерь алгоритма sAdaHedge и Random Forest Regression в моменты времени t, когда он был активен.

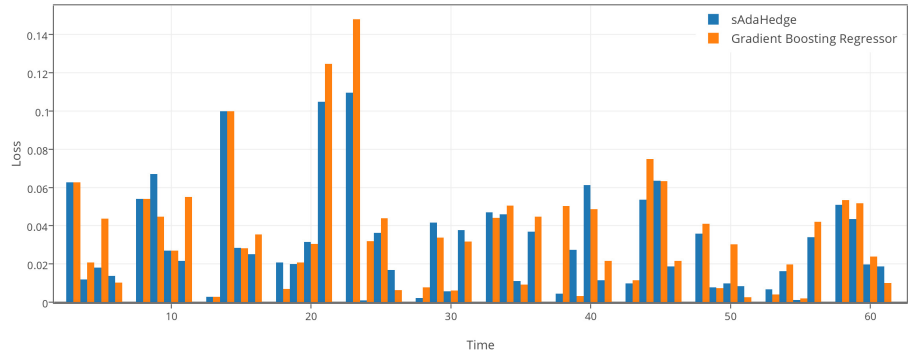

Рис. 8. Сравнение потерь алгоритма sAdaHedge и Gradient Boosting Regression в моменты времени t, когда он был активен.

Таблица 1. Итоговые кумулятивные потери.

sAdaHedge	**2.05**
Ridge Regressor	**2.94**
Lasso Regressor	**3.03**
Bayesian Ridge Regressor	**3.14**
Random Forest Regressor	**2.15**
Gradient Boosting Regressor	**2.32**

5. ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена задача агрегирования прогнозов специализированных экспертов. В качестве основного результата представлена модификация алгоритма AdaHedge из работы [11] для случая экспертов специалистов. Приведены теоретические обоснования и оценки того, что кумулятивные потери агрегирующего алгоритма не больше кумулятивных потерь специализированных экспертов в моменты времени, когда они были активными.

В алгоритме sAdaHedge мы приписываем спящим экспертам смешанные потери агрегирующего алгоритма, которые вычисляются по специальной формуле через потери активных на шаге t экспертов. Поэтому для анализ алгоритма sAdaHedge подобен анализу алгоритма AdaHedge из работы [11], с тем исключением, что вместо весов всех экспертов используются новые веса активных на данном шаге экспертов. Важный момент анализа в том, что смешанные потери m_t имеют представление как через веса всех экспертов так и через веса активных экспертов.

Был построен программный модуль для реализации задачи прогнозирования товарооборота крупной немецкой сети – ритейлера. Результат численного моделирования показывает, что метод вычисления агрегированных прогнозов со специалистами экспертами дает хорошие результаты для практического применения. А сам метод может применяться не только для прогнозирования товарооборота, но и в предсказании в других прикладных областях.

СПИСОК ЛИТЕРАТУРЫ

1. V'yugin V.V. The Following the Perturbed Leader Algorithm and Its Application for Constructing Game Strategies // Journal of Communications Technology and Electronics. 2015, 60(6), 647–657.

2. Vovk, V. A game of prediction with expert advice // Journal of Computer and System Sciences. 1998, 56, 153–173.

3. Vovk, V. Competitive on-line statistics // International Statistical Review. 2001, 69(2), 213–248.

4. Freund, Y., Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997, 55, 119–139.

5. Cesa-Bianchi, N., Lugosi, G. Prediction, Learning, and Games. Cambridge University Press. 2006.

6. В.В. Вьюгин Математические основы машинного обучения и прогнозирования. М.: Изд. МЦНМО, 304с. 2013.

7. Devaine, M., Gaillard, P., Goude, Y., Stoltz, G. Forecasting electricity consumption by aggregating specialized experts // Machine Learning. 2013, 90 (2), 231–260.

8. Kalnishkan, Y. Adamskiy, D. Chernov, A. Scarfe, T. Specialist Experts for Prediction with Side Information // IEEE International Conference on Data Mining Workshop (ICDMW). IEEE, 2015, 1470–1477.

9. Freund, Y., Schapire, R.E., Singer, Y., Warmuth, M.K. Using and combining predictors that specialize. In: Proc. 29th Annual ACM Symposium on Theory of Computing. 1997, 334–343.

10. Adamskiy, D., Koolen, W.M., Chernov, A., Vovk, V. A closer look at adaptive regret // Lecture Notes in Artificial Intelligence. 2012, 290–304.

11. S. de Rooij, T. van Erven, Grunwald, D., Koolen , M. Follow the Leader If You Can, Hedge If You Must // Journal of Machine Learning Research. 2014, 15, 1281–1316.

12. Blum, A. Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain // Machine Learning. 1997. 26, 5–23.

13. Bishop, C. Pattern Recognition and Machine Learning. Springer, New York. 2006.

Retail Advanced Demand Planning Using Aggregation of Specialized Prediction Strategies

Vladimir V. V'yugin, Azat I. Shamsutdinov

Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow, Russia
e-mail: vyugin@iitp.ru

Abstract: In the framework of prediction with expert advice a problem of aggregation of specialized experts is considered. The experts are algorithms that predict a sequence of outcomes online. We consider specialized experts – at each round only some of the experts output a prediction while the other ones are inactive. We present the optimal algorithm that aggregates these partial forecasts in one prediction. The property of asymptotic optimality is proved: at each round cumulative losses suffering by this algorithm do not exceed (up to some regret) cumulative losses of any other prediction strategy. An upper bound of regret (prediction error guarantee) is obtained. In our setting, these guarantees are not linked in any sense to a stochastic model: in fact, they hold for all sequences of consumptions, in a worst-case sense. We apply these rules to turnover retail network forecasting. The results of numerical experiments are presented.

Keywords: on-line prediction, prediction with expert advice, aggregation algorithm, sleeping experts, adaptive regret, demand planning.