УДК 519.2:004.9

И.Э. Зинькевич, Л.О. Кириченко, Т.А.Радивилова

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ПРОГНОЗИРОВАНИЯ СЛАБО КОРРЕЛИРОВАННЫХ ВРЕМЕННЫХ РЯДОВ

Аннотация. В работе рассмотрено прогнозирование слабо коррелированных временных рядов методами экспоненциального сглаживания, нейронной сети и дерева решений на примере данных реального интернет-магазина. Рассмотрены преимущества и недостатки каждого метода.

Ключевые слова: временной ряд, прогнозирование, экспоненциальное сглаживание, дерево принятия решений, сети долго-краткосрочной памяти.

Введение и цель

В современном мире, часто возникает потребность в анализе и предсказании временных рядов (ВР). ВР является являются распространенной и важной формой описания данных, так как позволяют наблюдать всю историю изменения интересующего нас значения. Это даёт нам возможность судить о «типичном» поведении величины и об отклонениях от такого поведения. Одной из сложной и интересной областей анализа ВР область электронной коммерции.

Электронная коммерция находится в постоянном развитии, чему способствуют новые технологии, услуги и тактические инструменты [1]. Чтобы «выжить» и выделиться среди множества интернет-магазинов, важно понимать поведение пользователя с момента первого прихода на сайт: отслеживать его перемещения, знать, какие продукты он посмотрел, положил в корзину, где кликал, что видел, в какой момент ушел, как и когда возвращался. В этом поможет веб-аналитика, которая подразумевает постоянный сбор, анализ и интерпретацию данных о посетителях, работу с основными метриками.

Качественная аналитика интернет магазина всегда начинается пути посетителя, который он прошел перед совершением покупки. Опишем условный путь пользователя, который пришел из соцсети:

посещение страницы; посетитель обращает внимание на пост; кликает на него; переходит на целевую страницу, на которую ведет ссылка из поста; изучает характеристики продукта; изучает способы оплаты и доставки; добавляет продукт в «Корзину»; оформляет заказ; совершает покупку.

На каждом из этапов пользователь может остановиться и закончить процесс, не совершив покупку. Процент конверсии высчитывается как отношение всех посетителей к количеству покупателей. Например, если сайт посетило 100 человек, но купило — 2, то конверсия равняется 2%. Данное измерение является основным в веб-аналитике всех коммерческих сайтов. Увеличение процента зависит от множества факторов: от дизайна страницы до ее функционала. Мониторинг конверсии позволяет вовремя понять, что электронный магазин нужно усовершенствовать.

Анализ и прогнозирование ВР дневных значений процента конверсии играет важнейшее значение для оптимизация эффективности онлайн-бизнеса. Однако, надо отметить, что практически все из классических методов анализа ВР базируются на вычислении корреляции между значениями ВР [2]. В случае слабо коррелируемых ВР, а также в случае, когда ВР имеет разряженную нулевыми значениями структуру, что характерно для многих сайтов электронных продаж, эти методы не подходят или имеют большую погрешность.

Широкое распространение для решения задач прогнозирования в последнее время получил нейросетевой подход. Нейронные сети позволяют моделировать сложные зависимости между данными в результате обучения на примерах. Однако прогнозирование ВР с помощью нейронных сетей имеет свои недостатки. Во-первых, для обучения нейронных сетей требуется ВР большой длины. Во-вторых, результат существенно зависит от выбора архитектуры сети, а также входных и выходных данных. В-третьих, нейронные сети требуют предварительной подготовки данных, или препроцессинга. Препроцессинг является одним из ключевых элементов прогнозирования: качество прогноза нейросети может решающим образом зависеть от того, в каком виде представлена информация для ее обучения. Общей целью препроцессинга является повышение информативности входов и выходов. Обзор методов выбора входных переменных и препроцессинга содержится в [3, 4].

В последнее время для анализа закономерностей временного ряда все чаще стали применяться методы Data Mining и машинного обучения [5], предназначенные для обнаружения различных шаблонов во временном ряде. При этом особую ценность в обнаружении таких шаблонов имеют логические методы. Эти методы позволяют находить логические if-then правила. Они пригодны для анализа и прогнозирования как числовых, так и символьных последовательностей, и их результаты имеют прозрачную интерпретацию.

Целью представленной работы является проведение сравнительного анализа прогнозирования ВР, на базе классических методов прогнозирования и методов машинного обучения на примере данных реального интернет-магазина.

Методы исследования

Электронная коммерция находится в постоянном развитии, чему способствуют новые технологии, услуги и тактические инструменты. Регулярно изменяются поставщики, диапазон покупателей, спектр товаров, что приводит к быстрому устареванию информации. Поэтому методы, которые требуют достаточно больших массивов ВР, такие как, например, модели авторегрессии и скользящего среднего, работают плохо.

Методы экспоненциального сглаживания. В основу экспоненциального сглаживания (ЭС) заложена идея постоянного пересмотра прогнозных значений по мере поступления фактических. Модель ЭС присваивает экспоненциально убывающие веса наблюдениям по мере их старения. Таким образом, последние доступные наблюдения имеют большее влияние на прогнозное значение, чем старшие наблюдения.

Модель ЭС имеет вид

$$Z(t) = S(t) + \varepsilon_t, \tag{1}$$

$$S(t) = \alpha \cdot Z(t-1) + (1-\alpha) \cdot S(t-1),$$

где α – коэффициент сглаживания; $0 < \alpha < 1$; $Z(t)$ – прогнозируемый ВР; $S(t)$ – сглаженный ВР; начальные условия определяются как $S(1) = Z(0)$. В данной модели каждое последующее сглаженное значение $S(t)$ является взвешенным средним между предыдущим значе-

нием временного ряда $Z(t)$ и предыдущего сглаженного значения $S(t-1)$.

Методы машинного обучения – чрезвычайно широкая и динамически развивающаяся область исследований, использующая огромное число теоретических и практических методов. Одним из методов машинного обучения является метод дерева принятия решений. Дерево принятия решений – средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Подобные деревья решений широко используются в интеллектуальном анализе данных. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе.

Каждый лист представляет собой значение целевой переменной, измененной в ходе движения от корня по листу. Каждый внутренний узел соответствует одной из входных переменных. Дерево может быть также «изучено» разделением исходных наборов переменных на подмножества, основанные на тестировании значений атрибутов. Это процесс, который повторяется на каждом из полученных подмножеств. Рекурсия завершается тогда, когда подмножество в узле имеет те же значения целевой переменной, таким образом, оно не добавляет ценности для предсказаний. В интеллектуальном анализе данных, деревья решений могут быть использованы в качестве математических и вычислительных методов, чтобы помочь описать, классифицировать и обобщить набор данных, которые могут быть записаны следующим образом: $(x, Y) = (x_1, x_2, x_3, \ldots, x_k, Y)$. Зависимая переменная Y является целевой переменной, которую необходимо проанализировать, классифицировать и обобщить. Вектор x состоит из входных переменных x_1, x_2, x_3 и т. д., которые используются для выполнения этой задачи [5].

 ISSN 1562-9945

Сети долго-краткосрочной памяти (Long Short Term Memory, LSTM) – особый вид рекуррентных нейронных сетей, способных к обучению долгосрочным зависимостям. Они были предложены Хохрейтером и Шмидхубером и доработаны и популяризованы другими в последующей работе. Они работают невероятно хорошо на большом разнообразии проблем и в данный момент широко применяются [3, 4].

LSTM специально спроектированы таким образом, чтобы избежать проблемы долгосрочных зависимостей. Запоминать информацию на длительный период времени - это практически их поведение по-умолчанию, а не что-то такое, что они только пытаются сделать. Все рекуррентные нейронные сети имеют форму цепи повторяющих модулей (repeating module) нейронной сети. В стандартной рекуррентные нейронные сети эти повторяющие модули будут иметь очень простую структуру.

Ошибки прогноза. Для получения количественных характеристик сравнительного анализа моделей были выбраны следующие характеристики ошибок прогноза. Среднее абсолютное отклонение (Mean Absolute Derivation, MAD) измеряет точность прогноза, усредняя величины ошибок прогноза. Использование MAD наиболее полезно в тех случаях, когда аналитику необходимо измерить ошибку прогноза в тех же единицах, что и исходный ряд. Эту ошибку вычисляют следующим образом:

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|X(t) - \hat{X}(t)\right|.$$

Среднеквадратическая ошибка (Mean Squared Error, MSE) – это другой способ оценки метода прогнозирования. Поскольку каждое значение отклонения возводится в квадрат, то этот метод подчеркивает большие ошибки прогноза. Ошибка MSE вычисляется следующим образом:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(X(t) - \hat{X}(t))^2.$$

Средняя абсолютная ошибка в процентах (Mean Absolute Percentage Error, MAPE) вычисляется путем отыскания абсолютной ошибки в каждый момент времени и деление ее на действительное наблюдаемое значение с последующим усреднением полученных абсолютных процентных ошибок. Этот подход полезен в том случае, когда

размер или значение прогнозируемой величины важны в оценке точности прогноза. MAPE подчеркивает насколько велики ошибки прогноза в сравнении с действительными значениями ряда. Данная ошибка вычисляется следующим образом:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|X(t) - \hat{X}(t)\right|}{X(t)}.$$

Входные данные

Входными данными в работе служили ежедневные данные сайта онлайн продаж, которые включали в себя количество кликов на сайт из социальных сетей, количество покупок и соответствующий коэффициент конверсии. Кроме этого имелась информация, какой язык использовал покупатель, из какой страны был сделан заказ и другие сведения.

На рис.1 представлены типичные ВР кликов, заказов и процентов конверсии. Ряды заказов и, соответственно конверсии, характеризуются нулевыми значениями, что значительно усложняет прогнозирование на следующий день. Корреляционная функция для ряда конверсии приведена на ррис.1 справа. Очевидно практически полное отсутствие корреляции между значениями ВР.

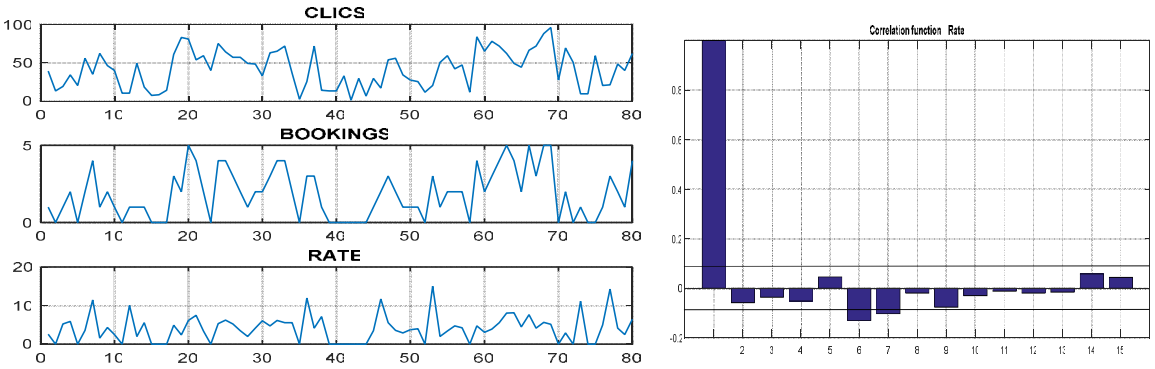

Рисунок 1 – Временные ряды кликов, заказов и процентов конверсии (слева); корреляционная функция для ряда процентов конверсии (справа)

Результаты исследования

Для построения моделей и нейронной сети использовался язык Python с библиотеками, реализующими методы машинного обучения. Для проведения прогнозирования ВР были разделены на две части, где первая использовалась для обучения модели, а вторая – для

оценки ее правдоподобности. Обучение моделей проводилось по S последним значениям (S было выбрано равным 20). Проверка моделей на прогнозирования m значений проводилась следующим образом: возьмем 20 последних значений первого ряда и сделаем прогноз на одно значений вперед, далее возьмем сдвинем наше окно на одно значение вперед, включив в окно прогноз для нового значения, и сделаем прогноз на одно значение вперед еще раз, и так m раз. На рис. 2 представлены результаты прогнозов каждой модели на 7 значений вперед. Сплошной линией показаны реальные значения, линия 1 – значения, полученные методом экспоненциального сглаживания, 2 – на основе дерева решений, 3 – с помощью нейронной сети.

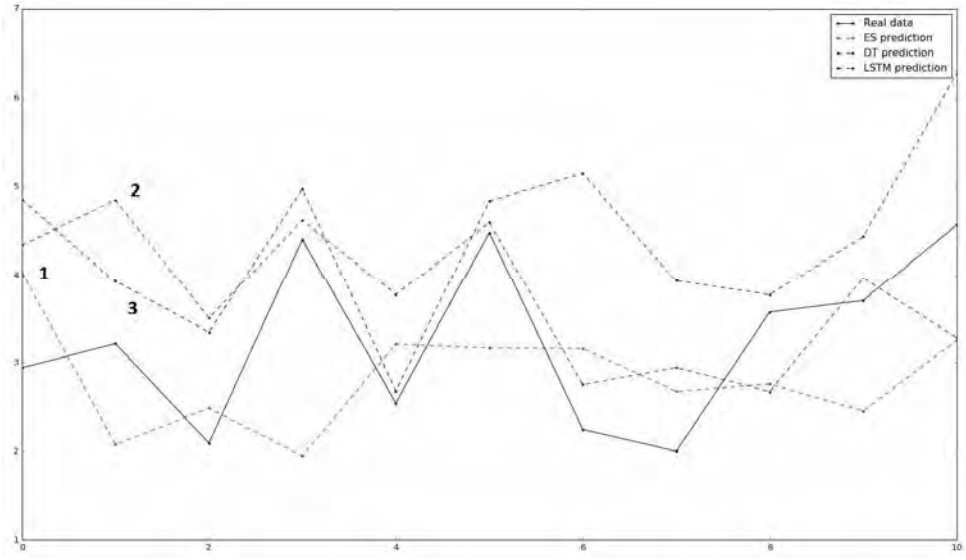

Рисунок 2 – Прогнозные значения для разных моделей

Были рассчитаны прогнозные значения для S =20 и m=1 (такой выбор параметров определяется требованиями интернет-магазина) по 100 значениям ряда для множества ВР процента конверсии. Результаты вычислений, характерные для большинства рядов, приведены в таблице 1.

Таблица 1

Погрешности методов

Метод	Экспоненциальное сглаживание	Дерево решений	Нейронная сеть
MAD	0.013751	0.014218	0.002645
MSE	0.000369	0.000353	0.000012
MAPE	0.491	0.513	0.072

В результате анализа прогнозов разных значений S и m было установлено, что метод экспоненциального сглаживания, не смотря на свою простоту и не требовательность в количестве данных, по которым будет построен прогноз, имеет в большинстве случаев наименьшие погрешности прогнозируемых значений, но в то же время, некоторые прогнозные значения значительно удалены от реальных. Метод дерева решений показал себя неудобным в выборе параметров и имеющим погрешности, сопоставимые с ошибками экспоненциального сглаживания, но без сильно удаленных прогнозных значений. Нейронная сеть LSTM, которая имеет более сложную структуру и её необходимо предварительно обучить на достаточно большом временном ряде, показала отличные результаты, как и в общей погрешности прогнозов, так и в удаленности прогнозов от реальных значений временного ряда.

Выводы

Результаты исследования методов прогнозирования слабо коррелированных временных рядов, типичных для рядов конверсии в электронной коммерции, показали, что экспоненциальное сглаживание является самым простым, быстрым и удобным в настройке методом прогнозирования, однако в случае сложных или долгосрочных зависимостей становится не применим. Метод дерева решений быстрый в обучении, не сложен для понимания, но неудобен в выборе параметров и плохо работает при обучении на данных, которые имеют много признаков. Нейронная сеть является громоздкой, долгой в обучении, требует множества параметров, которые нужно подбирать, но имеет очень хорошие показатели в прогнозировании и на порядок меньшие ошибки.

ЛИТЕРАТУРА

1. [Электронный ресурс]:
http: // lpgenerator.ru/blog/2015/07/02/kakoj-dolzhna-byt-veb-analitika-internet-magazina
2. Ханк Д. Бизнес-прогнозирование / Ханк Д. Изд. Дом «Вильямс», 2003. - 656 с.
3. Guyon I. An Introduction to Variable and Feature Selection. J. Guyon Isabelle, Elisseeff Andre / Of Machine Learning Research 2003. –С.1157-1182.
4. Ежов А.А Нейрокомпьютинг и его применения в экономике и бизнесе / Ежов А.А., Шумский С.А. –М., 1998. –С.216.
5. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. / Вьюгин В.В.–М., 2013. –С.387.