УДК 330.1
doi:10.18720/SPBPU/2/id24-519

Мохаммад Хани [1],
аспирант;
Пак Вадим Геннадьевич [2],
доцент, канд. физ.-мат. наук

ОБНАРУЖЕНИЕ СОБЫТИЙ НА ОСНОВЕ АНАЛИЗА СВЯЗАННЫХ В ПРОСТРАНСТВЕ И ВРЕМЕНИ ДАННЫХ

[1,2] Россия, Санкт-Петербург,
Санкт-Петербургский политехнический университет Петра Великого,
[1] mohammad.h@edu.spbstu.ru, [2] pak_vg@spbstu.ru

Аннотация. Статья посвящена теме обнаружения событий на основе анализа данных, связанных с пространством и временем. В настоящее время изучается роль использования нейронных сетей в обработке цифровых данных, которые могут быть получены от компаний мобильной связи, с целью выявления социальной активности, происходящей где-либо в модели умного города. Свойства коммуникационных данных объясняются их привязкой ко времени и месту, что позволяет прогнозировать вероятность социальных событий, происходящих в данном месте и в данное время. В последнее время глубокое обучение стало значительно более прогностичным. Глубинные модели использовались во многих исследованиях для обнаружения аномалий; большинство из них основаны на нейронной сети LSTM без учета пространственных особенностей. Или на сверточных нейронных сетях (CNN). Ни в одном из предыдущих исследований нейронная сеть на основе ConvLSTM не применялась к этим данным. Использование ConvLSTM в большинстве исследований ограничивалось видеоданными.

Ключевые слова: пространственно-временные данные, ConvLSTM, расстояние Махаланобиса, обнаружение событий.

Mohammad Hani [1],
Postgraduate Student;
Vadim G. Pak [2],
Associate Professor, Candidate of Physical and Mathematical Sciences

EVENT DETECTION BASED ON THE ANALYSIS OF SPATIO-TEMPORAL DATA

[1,2] Peter the Great St.Petersburg Polytechnic University, St. Petersburg, Russia,
[1] mohammad.h@edu.spbstu.ru, [2] pak_vg@spbstu.ru

Abstract. The article is devoted to the topic of event detection based on the analysis of data related to space and time. The role of using neural networks in processing digital data that can be obtained from mobile communications companies is being studied in order to discover the social activities happening somewhere in the smart city model. The properties of communication data are explained by their connection with time and place, which makes it possible to predict the possibility of social events occurring in a given place and time. Recently, deep learning has become significantly more predictive. Deep models have been used in many studies to detect anomalies; Most of them are based on an LSTM neural network without taking into account spatial features. Or based on convolutional neural networks (CNN). No previous research has applied a ConvLSTM-based neural network to this data. The use of ConvLSTM in most studies was limited to video data.

Keywords: spatio-temporal data, ConvLSTM, Mahalanobis distance, event detection.

Введение

В век технологий и информации данные и информация являются основным и вдохновляющим источником для властей и советов умных городов при планировании предоставления услуг и удобств своим гражданам и обеспечении их безопасности.

С. Джеффри разработали модель прогнозирования сотового трафика данных, основанную на глубоком обучении. Вначале данные собирались по времени и брали одну характеристику – интернет-сессии, а затем он использовал нейронную сеть LSTM, и нейронную сеть прямого распространения (FFNN) и сравните результат прогнозирования с базовой моделью ARIMA и нейронной сетью прямого распространения (FFNN). Результаты показали, что модель, основанная на сети LSTM, дала лучшие результаты, чем две другие модели, и имела более короткое время обучения, чем FFNN. Она также показала лучшую производительность для ARIMA по сравнению с FFNN [2].

С. ЭльЭлими и С. Мустафа применил алгоритм ARIMA (2, 1, 0) и нейронную сеть LSTM, а модели были протестированы на трех разных квадратах в природе, и производительность этих моделей была неодинаковой для разных квадратов. Они анализировали данные с течением времени еженедельно и ежечасно, используя функцию «только Интернет» для проверки прогноза, и результаты были хорошими для обеих моделей [3].

М. С. Парвез сосредоточился на поиске аномалий в данных, полагаясь на алгоритмы кластеризации, где он проверял k-средние; Алгоритмы иерархической кластеризации основаны на данных заданного квадрата для суммы четырех атрибутов: входящие и исходящие вызовы и входящие или исходящие текстовые сообщения в качестве нового атрибута. По окончании кластеризации кластером с наименьшим количеством элементов считался кластер, содержащий аномалию [4].

Сверточная Long Short-Term Memory-нейронная сеть

C. Xiaoa, N. Chen, C. Hu, K. Wang, Z. Xu, Y. Cai, разрабатывают новую модель глубокого обучения.

Архитектура ConvLSTM сочетает в себе возможности CNN и LSTM. ConvLSTM обычно разрабатывается для многомерных пространственно-временных данных, таких как спутниковые изображения. Предыдущие уравнения LSTM были изменены для описания компонентов ConvLSTM, добавив в вентили операции произведения свертки [1].

Математическая модель ConvLSTM:

$$
\begin{aligned}
c_t' &= \sigma_c(W_{xc} \otimes x_t + W_{hc} \otimes h_{t-1} + b_c),\\
i_t &= \sigma_g(W_{xi} \otimes x_t + W_{hi} \otimes h_{t-1} + W_{ci} \odot c_{t-1} + b_i),\\
u_t &= \sigma_g(W_{xu} \otimes x_t + W_{hu} \otimes h_{t-1} + W_{cu} \odot c_{t-1} + b_u),\\
o_t &= \sigma_c(W_{xo} \otimes x_t + W_{ho} \otimes h_{t-1} + W_{co} \odot c_t + b_o),\\
c_t &= u_t \odot c_{t-1} + i_t \odot g_t,\\
h_t &= o_t \odot \sigma_c(c_t),
\end{aligned}
$$

где пояснения представлены в табл. 1.

Таблица 1

Описание показателей, входящих в уравнение модели ConvLstm

c_t	Состояние клетки в момент времени t	x_t	Цепочка доходов единицы LSTM
h_t	Скрытое состояние на временном шаге t	$W_{xo,xc,xi,xu}$	Ядро свертки применяется к трехмерному входному тензору x_t в каждом компоненте
i_t	Выходной сигнал входного затвора на временном шаге t	$W_{ho,hc,hi,hu}$	Ядро свертки, примененное к трехмерному входному тензору h_t в каждом компоненте
u_t	Забыть выход ворот на временном шаге t	$b_{o,j,i,u}$	Коэффициенты смещения для каждого компонента
g_t	ячейка-кандидат на временном шаге t	σ_c	Продолжайте активировать портал (сигмоид)
o_t	Выход выходного затвора находится на временном шаге t	σ_g	Продолжайте активировать портал (танх)
$\odot$	Продукт Адамара	$\otimes$	сложить (сложить) изделие

Расстояние Махаланобис

Это эффективная многомерная мера расстояния, которая измеряет расстояние точки (наблюдения) от распределения данных. Эта шкала была введена профессором П. К. Махаланобисом в 1936 г.

Х. Горбани рассказывает об использовании расстояния Махаланобиса, если признаки в наборе данных коррелируют друг с другом, значения ковариационной матрицы будут большими; таким образом, деление на большую ковариацию (умножение на обратную ковариационную матрицу) эффективно уменьшит расстояние. Напротив, если переменные не коррелируют, ковариация мала и расстояние существенно не уменьшается. Таким образом, он решает, как проблемы стандартизации, так и взаимозависимости переменных, которые не может решить евклидово расстояние. При расчете евклидовых расстояний расстояние между p2 и ближайшей точкой больше, чем расстояние между p1 и ближайшим соседом. При использовании шкалы расстояний Махаланобиса мы обнаруживаем, что эти два расстояния равны. [5]

1. Постановка задачи

1.1. Описание предметной области

Процесс анализа данных проходит в несколько этапов, а именно: Сбор данных, Хранение данных, Обработка данных, Очистка данных, Анализ данных. Каждый из этих этапов считается важным этапом для следующего, поскольку точность результатов одного этапа влияет на точность следующего этапа.

Сбор данных: Данные, используемые в исследовании, представляют собой телекоммуникационные данные из открытых источников. В рамках конкурса Big Data Challenge в 2014 году, опубликованного Telecom Italia и SpazioDati, наборы коммуникационных данных SMS, звонков и Интернета были предоставлены для Милана и Торонто в Италии; пространственно связаны.

Хранение данных: Используемые данные представляют собой цифровые данные, хранящиеся в текстовых файлах.

Обработка данных. На этом этапе текстовые файлы считываются и преобразуются в структуры с использованием языка программирования Python.

Очистка данных. На этом этапе мы изучили недостающие данные и заменили их средним арифметическим значений для каждого типа информации. Полные данные, относящиеся к предназначенной для этого области. Исследования и территории вокруг нее также использовались с целью изучения влияния предполагаемой территории на данные из прилегающих территорий.

Анализ данных: это шаг, на котором необработанные данные преобразуются в полезную информацию.

1.2. Определение проблемы

– Как мы можем идентифицировать социальные события на основе обнаружения аномалий?

– Каковы методы обнаружения аномалий в неразмеченных данных?

– Как можно смоделировать пространственно-временные данные?

2. Моделирование системы

2.1. Обоснование выбора языка моделирования

2.1.1. Обработка исходных данных

В этой статье использовался набор данных по городу Милан, охватывающий 550 квадратных километров города Милана; Он был разделен на группу квадратов, у каждой коробки есть идентификатор.

Сбор данных основан на записях данных о вызовах (CDR). Эти журналы содержат различные атрибуты активности, отражающие активность пользователя, а именно:

– SmsIn: представляет значение, пропорциональное количеству SMS, полученных ящиком за определенный период времени.

– SmsOut: представляет собой значение, пропорциональное количеству SMS-сообщений, отправленных ящиком за определенный период времени.

– CallIn: представляет значение, пропорциональное количеству вызовов, полученных ящиком в течение определенного периода времени.

– CallOut: представляет значение, пропорциональное количеству вызовов, отправленных ящиком в течение определенного периода времени.

– Интернет: количество записей, созданных для начала или завершения подключения к Интернету в пределах квадрата во временной области. Создается, когда сеанс длится 15 минут или каждые 5 МБ израсходованных.

Если в пространственном блоке не происходит никаких действий, для этого блока не записывается никакая запись. Данные были собраны для каждого квадранта для каждого временного интервала, а также собраны репрезентативные значения для звонков, сообщений и интернет-сессий. Мы отмечаем, что в соответствии с политикой конфиденциальности данных значения атрибутов представляют собой не реальные значения, а скорее значения, пропорциональные реальным значениям. Более высокое значение представляет собой наибольшую активность для каждого атрибута данных.

На рисунке 1 показан образец набора данных города Милан. Первый столбец – это идентификатор ящика.

	Time	Id	SmsIn	SmsOut	CallIn	CallOut	Internet
1439981	1383432600000	9996	0.168332	0.580365	0.316518	0.0	36.581953
1439982	1383432600000	9997	0.171990	0.720503	0.243143	0.0	38.845497
1439983	1383432600000	9998	0.171990	0.693185	0.256862	0.0	38.414371
1439984	1383432600000	9999	0.186697	0.525312	0.245184	0.0	27.121305
1439985	1383432600000	10000	0.226221	0.638219	0.145000	0.0	24.142753

Рис. 1. Характеристики коммуникативной деятельности

Если в течение определенного периода времени внутри ящика не происходит никакой активности, журнал не создается. В то время как значение NAN для атрибута представляет отсутствие какой-либо активности во временной области. Короче говоря, эти данные содержат временную информацию, представленную десятиминутным интервалом времени, в дополнение к пространственным характеристикам, представленным идентификатором географического квадрата. Помимо активности коммуникаций, движение во времени и пространстве.

Поскольку данные не размечены, в качестве целевого мы рассматриваем значения атрибутов активности пользователей в следующий момент времени.

На этом этапе были собраны данные для территорий, окружающих район Сан-Сиро в Милане, простирающихся на 16 квадратов в длину и 16 квадратов в ширину.

Мы собрали значения входящих и исходящих атрибутов предыдущих коммуникаций за период 20 минут, значения предыдущих входящих и исходящих сообщений за период времени 20 минут, и пользоваться Интернетом сроком на 20 минут.

2.1.2. Статистика по утвержденным данным

Мы можем определить временной ряд как вектор X такой, что:

$$X = \{x_1, x_2, x_3, x_4 \dots, x_t\},$$

где x_t представляет данные в момент времени $i \in T$,T ={1,2,…….,t}.

Изучая данные, мы пришли к следующим наблюдениям:

– активность города варьируется утром, в полдень и вечером;

– активность города варьируется между рабочими днями и официальными еженедельными днями отдыха;

– существует взаимосвязь между функциями данных (входящие и исходящие сообщения, отправленные и полученные сообщения и использование Интернета).

Данные, подлежащие изучению, должны быть стабильными и предсказуемыми, чтобы исключить любую явную корреляцию и кол линеарность с предыдущими данными.

Расширенный тест Дики–Фуллера (ADF) проверяет нулевую гипотезу о наличии единичного корня в выборочном временном ряду. Это версия теста Дики–Фуллера, но для более крупного и сложного набора временных рядов. Статистика ADF, используемая в тесте, представляет собой отрицательное число. Чем более отрицательное значение, тем больше отвергается гипотеза, доказывающая существование авторегрессии.

Процедура тестирования для теста ADF такая же, как и для теста Дики–Фуллера, но она применяется к модели.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

где α – константа, β – коэффициент временного тренда, а p – порядок запаздывания авто регрессионного процесса.

Таблица 2

Результат теста ADF в ящике 5638 для пяти атрибутов

Поле данных	Статистика тестирования	p-value	Стационарный
SmsIn	−15,112500	7,660685176721849e-28	Да
SmsOut	−13,772560	9,621714586151542e-26	Да
CallIn	−14,049726	3,1934266203633457e-26	Да
CallOut	−13,876625	6,321157964243545e-26	Да
Internet	−12,193763	1,2682583594087165e-22	Да

2.2. Построение модели

На этапе проектирования модели, поскольку основной целью является использование временной и пространственной информации, имеет смысл установить связи между соседними квадратами в миланской сетке, а также временные связи.

Отсюда и возникла идея использования модели глубокого обучения на основе нейронной сети ConvLSTM, где функция Convolution находит пространственные корреляции между квадратами, а LSTM отличается способностью находить временные корреляции.

Модель, состоящая в основном из слоя ConvLSTM.

Слои Reshape и Permute – это первые шаги для инициализации соответствующих входных данных для слоя ConvLSTM формы (TimeSteps_Samples, Height, Width, Channels).

Слой шума погружения добавляет шум к входным данным. Это типичный процесс, который помогает предотвратить переобучение нейронной сети обучающими данными.

Использовался слой ConvLSTM с количеством фильтров, равным десяти, и ядром свертки размером 5x5; с ReLU в качестве функции активации. При использовании функции итеративной активации по умолчанию.

В дополнение к функции отсева 20 % для весов, чтобы избежать переобучения данных обучения, он отключает нейроны (или делает некоторые веса равными нулю) случайным образом, чтобы сеть не полагалась на характеристики обучающих данных и могла работать с тестовыми данными, которые значительно отличаются от обучающих данных. Слой Conv2D для получения желаемой выходной формы с ядром 5x5.

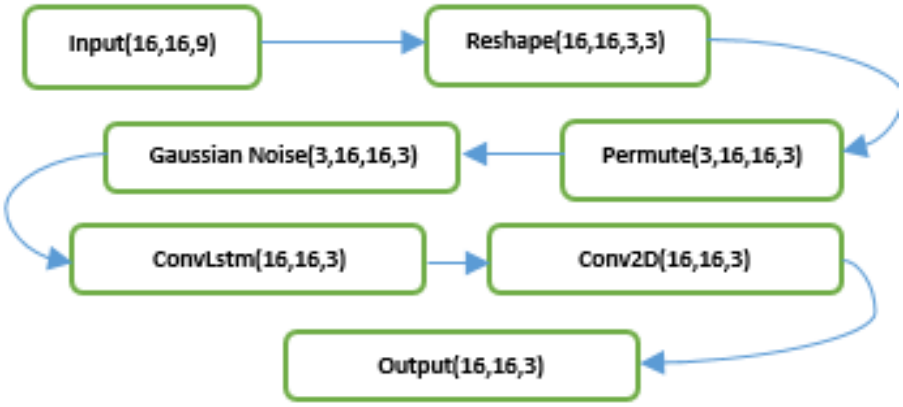

Рис. 2. Архитектура базовой модели на основе ConvLSTM

2.2.1. Обнаружение событий

Принцип, принятый для различения выбросов и социальных событий, заключается в том, что социальные события длятся более 15 минут.

Для каждого деления каждого квадрата выполняем описанные ниже действия.

Моделирование распределения данных в рамках нормального (ныряющего) распределения на основании того, что оценку максимального правдоподобия можно рассчитать для нормального распределения данных $\boldsymbol{x_1, \ldots, x_N}$.

Однако поскольку эта оценка очень чувствительна к наличию выбросов (событий) в наборе данных, соответствующие расстояния Махаланобиса вводят в заблуждение и являются неточными. Поэтому мы использовали робастную ковариационную оценку, в которой мы полагались на использование метода MCD для более точной оценки ковариационной матрицы.

В зависимости от расчета как $\hat{\mu}$, так и ковариационной матрицы $\hat{\sum}$; Мы рассчитываем расстояние Махаланобиса для каждого наблюдения по данным.

На этапе обнаружения аномалий мы полагаемся на ошибку прогноза; Мы сравниваем прогнозируемую сеть с реальной сетью и вычисляем разницу между соответствующими квадратами. Таким образом, мы вычисляем аномальные значения в рамках распределения разностей данных между реальными значениями и ожидаемыми значениями.

Затем мы находим точки с наибольшим расстоянием по разнице распределения пяти атрибутов на основе расстояния Махаланоби и критерия хи-квадрат, которые являются кандидатами на представление социальных событий в городе.

2.2.2. Тестирование и внедрение

Данные были разделены на две группы.

1) Группа обучения и проверки, которая действует с 1 ноября 2013 г. по 26.11.2013 г., где 100 % было отведено обучению.

2) Тестовая группа продлится с 26.11.2013 по 01.01.2014, где 30 % на валидацию и 70 % на тестирование.

Эксперимент проводился с использованием нескольких оптимизаторов (Adam, Adadelta, SGD, Adagrad, RMSProp, Adamax) и функций потерь (Binary_crossentropy, MSE, MAE) (см. табл. 3).

Функция ошибок Binary_crossentropy не дала хороших результатов. При этом функция потерь MSE дала наилучшие результаты.

В результате в декабре в районе стадиона Сан-Сиро было зафиксировано 5 инцидентов из 6 (см. табл. 4).

Таблица 3

Оценка производительности предлагаемой модели по используемой функции оптимизации

Оптимизатор	Функция потерь	Тестирование	Точность
Adam	MAE	0,0465	0,7568
Adam	MSE	0,0052	0,7439
Adam	binary_crossentropy	0,4811	0,6844

Таблица 4

Матчи

	Зрители	**Матч**	**Время**	**Лига**	**День недели**
1	43 706	Internazionale vs. Sampdoria 1-1	1 декабря 2013 15:00	Inter Milan season	Будни
2	12 714	Internazionale vs Trapani 3–2	4 декабря 2013 21:00	Coppa Italia	Будни
3	33 732	Internazionale vs. Parma 3-3	8 декабря 2013 20:45	Inter Milan season	Выходной
4	61 744	Milan vs. Ajax 0-0	11 декабря 2013 20:45	Inter Milan season	Будни
5	79 311	Internazionale vs. Milan 1-0	22 декабря 2013 20:45	A.C. Milan season	Выходной

И 16 декабря 2013 г .Кристиан Сапата, левый игрок «Милана», вместе со своими товарищами по команде празднует забитый гол в матче итальянской лиги против «Ромы»

Заключение

В результате мы обнаружили эффективность ConvLSTM в обнаружении событий и попытке извлечения пространственной и временной информации, но мы ожидаем в ближайшем будущем обработки других источников данных для той же исследуемой области и в тот же период

времени, и использование ConvLSTM в генеративно-состязательной сети с целью достижения точного прогнозирования целевых данных.

Список литературы

1. Xiaoa C., Chen N., Hu C., Wang K., Xu Z., Cai Y. [et al.] A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data // In: Environmental Modelling & Software. – ELSEVIER, 2019.

2. Jaffry S. Cellular traffic prediction with recurrent neural network // arXiv.org. – 2020. – arXiv:2003.02807.

3. ElElimy S., Moustafa S. Big Data in telecom industry: effective predictive techniques on CDRs // SC 20(11): e1. – EAI, 2020. – DOI: 10.4108/eai.13-7-2018.164919.

4. Parwez M. S., Rawat D. B., Garuba M. Big Data analytics for user-activity analysis and user-anomaly detection in mobile wireless network // IEEE Transactions on Industrial Informatics. – 2017. – P. 12.

5. Ghorbani H. Mahalanobis distance and its application for detecting multivariate outliers // Facta Univ Ser Math Inform. – 2019. – Vol. 34. – Pp. 583–595. – DOI:10.22190/FUMI1903583G.

6. Dataset [Electronic resource] / The Telecom Italia. – URL: https://dandelion.eu/datagems/SpazioDati/telecom-sms-call-internet-mi/description/ (access date: 10.06.2024).