ПРОГНОЗИРОВАНИЕ МОБИЛЬНОГО ТРАФИКА ИЗ НЕОБРАБОТАННЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ СЕТЕЙ LSTM

В.Е. Митрохин, И.Н. Башков

Омский государственный университет путей сообщения, Омск, Российская Федерация

Рассмотрен методологический подход к прогнозированию сетевого трафика в мобильных сетях, в частности в сетях LTE. Прогнозирование сетевого трафика жизненно необходимо в современных высоконагруженных сетях, а также в сетях с динамическим перераспределением ресурсов. Представлен и обоснован подход к прогнозированию, заключающийся в использовании рекуррентных нейронных сетей, а также исследовано влияние на достоверность модели ее компонентов. Результаты в литературе показывают, что сети LSTM превосходят другие подходы машинного обучения для анализа временных рядов в прогнозировании трафика.

Ключевые слова: прогнозирование трафика, мобильный трафик, LTE, LSTM, краткосрочная и долгосрочная память.

MOBILE TRAFFIC PREDICTION FROM RAW DATA USING LSTM NETWORKS

V.E. Mitrokhin, I.N. Bashkov

Omsk State Transport University, Omsk, Russian Federation

A methodological approach to the prediction of network traffic in mobile networks, in particular in LTE networks, is considered. Predicting network traffic is vital in today's highly loaded networks, as well as in networks with dynamic resource redistribution. The predicting approach, which consists in the use of recurrent neural networks, is presented and substantiated, and the effect on the reliability of the model of its components is investigated. The results in the literature show that LSTM networks are superior to other machine learning approaches for time series analysis in traffic prediction.

Keywords: prediction, mobile traffic, LTE, LSTM, long and short term memory.

Прогнозный анализ трафика мобильной сети приобретает фундаментальное значение для сотовой сети следующего поколения. Упреждающее знание требований пользователя позволяет системе оптимально распределять ресурсы. В данной работе мы изучаем мобильный трафик базовой станции LTE и разрабатываем систему прогнозирования трафика с использованием рекуррентных нейронных сетей. Информация о мобильном трафике собирается из физического нисходящего канала управления (PDCCH) LTE с помощью пассивного инструмента, представленного в [1]. С помощью этого инструмента мы можем собирать всю управляющую информацию с разрешением 1 мс от базовой станции. Эта информация содержит блоки ресурсов, размер транспортного блока и схему модуляции, назначенную каждому пользователю, подключенному к eNodeB. Конструкция системы прогнозирования включает в себя блоки долговременной кратковременной памяти. Что касается многоуровневых сетей перцептронов или других структур нейронных сетей, рекуррентные сети выгодны для задач с последовательными данными (например, языковое моделирование) [2]. В нашем случае мы ставим задачу контролируемого многомерного предсказания мобильного трафика, где целью является минимизация ошибки предсказания с учетом информации, извлеченной из PDCCH. Мы оцениваем одношаговое предсказание и ошибки долгосрочного предсказания предложенной методологии, рассматривая различные числа для продолжительности наблюдаемых значений, которые определяют длину памяти сети LSTM и сколько информации должно быть сохранено для точного предсказания трафика.

Результаты в литературе показывают, что сети LSTM превосходят другие подходы машинного обучения для анализа временных рядов в прогнозировании трафика. Структуры LSTM были предложены в [3]. Здесь наборы данных состоят из пространственно-временного распределения мобильного трафика в различных базовых станциях. Пространственная корреляция была использована для доказательства сходства между соседними базовыми станциями. Несмотря на то, что LSTM был применен для прогнозирования трафика, входные данные учитывают только одну метрику (например, трафик пространственно распределенный), и рассматривается только одношаговое прогнозирование.

Используя методологию, как это сделано в [1] и [4], мы можем собирать информацию о планировании LTE пользователей, подключенных к определенному eNodeB. Преимуществом этого инструмента является большой объем информации и временная детализация данных. Мы можем декодировать сообщения управляющей информации нисходящей линии связи (DCI) от PDCCH, которые содержат схему модуляции и кодирования (MCS), количество ресурсных блоков и размер транспортного блока, назначенного пользователям в каждой миллисекунде. Эта информация доступна как для направления нисходящей линии связи, так и для направления восходящей линии связи. Анонимность данных обеспечивается тем, что пользователи идентифицируются ограниченным числом временных идентификаторов радиосети (RNTI), которые обновляются после 10...15 секунд бездействия. Доступна только управляющая информация планирования: PDCCH не зашифрован, в то время как нисходящие и восходящие фактические данные передаются по зашифрованным физическим каналам для обеспечения конфиденциальности пользователей.

Набор данных состоит из информации за один месяц, собранной системой путем мониторинга различных eNodeB.

Пусть $D = \{D_{c1}, D_{c2}, \dots\}$ набор данных, где D_{ck} набор измерений для контролируемой ячейки k . Учитывая D_{ck} для каждого подключенного пользователя в момент времени t , временно идентифицированный RNTI r , мы декодируем сообщение DCI, содержащее блоки ресурсов, размер транспортного блока и другую запланированную информацию для направлений восходящей и нисходящей линии связи. Мы храним эту информацию в мере $s_r(t)$, где t однозначно соответствует номеру подкадра LTE или TTI, который составляет 1 мс.

Мы вычисляем совокупные измерения ячейки трафика для данного временного интервала T , который является суммой трафика, генерируемого всеми RNTI, подключенными в течение временного интервала T , $R(T)$

$$S(T) = \sum_{r(t) \in R(T)} \sum_{t \in T} S_r(t).$$

Таким образом, $S(T)$ – это вектор, который содержит количество ресурсных блоков, выделенных в восходящем и нисходящем направлениях, количество сообщений, отправленных в обоих направлениях, и общую сумму размеров транспортных блоков для заданного временного интервала T . Кроме того, для каждого временного интервала T мы включаем количество присоединенных пользователей к eNodeB во временном интервале T . В нашем случае мы рассматриваем T как количество TTI, для которых мы агрегируем трафик.

Рекуррентные нейронные сети являются обобщением нейронных сетей прямой связи, которые были разработаны для решения временных и прогнозных задач. LSTM – это особый вид RNN, которые были введены в [5]. Они были явно разработаны, чтобы избежать проблемы долгосрочной зависимости, которая является причиной проблемы исчезающего градиента в нормальных RNNs [6]. Возможность изучения долгосрочных зависимостей обусловлена структурой блоков LSTM, которая включает в себя элементы, регулирующие процесс обучения.

В стандартном блоке LSTM (рис. 2) основные операции выполняются входным затвором i_t , забывающим затвором f_t и выходным затвором o_t . Кроме того, состояние ячейки c_t представляет собой память устройства, и она обновляется информацией, которая должна храниться (или быть забыта), предоставленной входным затвором (или забывающим затвором).

Рис. 1. Нормализованный недельный трафик с ноды

Рис. 2. Стандартный LSTM блок

Блок LSTM объединяет выход предыдущего блока h_{t-1} с текущим входом x_t , используя вход, выход и элементы forget для обновления памяти ячейки. Переменные i_t и f_t представляют, соответственно, информацию, которую необходимо сохранить или забыть из прошлого и текущего ввода. Состояние ячейки c_t обновляется путем суммирования предыдущего состояния ячейки c_{t-1} и состояния ячейки-кандидата $\tilde{c}_t$, взвешенного соответственно с помощью f_t и i_t . Наконец, мы получаем выход h_t , применяя функцию к c_t и умножая ее на o_t . Затем текущий выход h_t передается на следующий блок и объединяется с входом при следующем временном индексе $t + 1$.

Несколько блоков LSTM объединяются для формирования одного уровня сети LSTM (рис. 3). Каждый блок вычисляет операции над одним временным индексом и передает выходные данные на следующий блок LSTM. Число соединенных ячеек указывает количество наблюдений данных, которые рассматриваются перед выполнением прогноза. В нашем случае входным сигналом является вектор трафика eNodeB $S(T)$, а число наблюдений – число выбранных временных интервалов T .

Предлагаемая архитектура для прогнозирования мобильного трафика изображена на рис. 4. Мы рассматриваем несколько уровней базовых блоков LSTM для формирования многоуровневой сети LSTM. Многоуровневая LSTM сеть способна изучать совокупность временные зависимости мобильного трафика: блок LSTM каждого уровня извлекает фиксированное количество объектов, которые передаются на следующий уровень. Глубина сети (например, количество уровней) должна увеличивать точность предсказания, которое выполняется последним, полностью связанным слоем. Для одношагового прогнозирования мы используем архитектуру «все к одному», что означает, что сеть анализирует мобильный трафик в течение фиксированного количества временных интервалов до T и затем пытается предсказать трафик в следующем временном интервале $T + 1$. В многошаговом предсказании мы задерживаем предсказание для выбранного числа временных шагов подобно тому, что делается для задач моделирования языка, когда они пытаются предсказать последовательность слов. Наконец, выходные данные из сети LSTM передается в подключенную нейронную сеть, которая выполняет фактическое предсказание. Задача для последнего уровня – это проблема регрессии, которую мы пытаемся решить: с точки зрения реализации этот уровень применяет функцию активации softmax, которая необходима во время фазы обучения для оптимизации весов сетевых элементов.

Рис. 3. Одноуровневая LSTM сеть

Рис. 4. Предлагаема архитектура для предсказания мобильного трафика

Мы используем набор данных мобильного трафика из двух различных eNodeBs, которые мы собрали в течение одного месяца, чтобы оценить производительность предлагаемой архитектуры. Для каждый eNodeB мы вычисляем совокупный трафик ячейки.

В качестве метрики для измерения точности алгоритма прогнозирования мы выбираем нормированную Среднеквадратическую ошибку (NRMSE), которая задается как

$$NRMSE = \frac{1}{\bar{x}} \sqrt{\frac{\sum_{t=1}^N (\tilde{x}_t - x_t)^2}{N}}$$

где N – общее число точек, $\tilde{x}_t$ и x_t прогнозируемые значения и соответствующее ему наблюдение в момент времени t , $\bar{x}$ – их средним значением. Эта же метрика используется для сравнения точности предложенной архитектуры с точностью, полученной с помощью других алгоритмов прогнозирования.

Реализация алгоритма прогнозирования мобильного трафика выполнена при помощи Python, используются библиотеки Keras и Tensorflow в качестве бэкэнда. Количество скрытых уровней равняется 5 – это один из гиперпараметров, которые необходимо выбрать и который влияет на компромисс между точностью предсказания и временем, необходимым для обучения сети. Большее число уровней может увеличить точность предсказания, но мы хотим сосредоточиться на соотношении между количеством прошлых наблюдаемых значений и точностью многошагового предсказания, которое определяет количество информации, необходимой для запоминания и использования сетью. По этой же причине мы устанавливаем количество промежутков равным 100. Данные за три недели используются для обучения и проверки архитектуры. Следующие результаты взяты за прошлую неделю. Мы используем метод оптимизации Adam [7] для итерационного обновления весов сети на основе обучающих данных.

Далее мы представляем результаты многошагового прогнозирования, то есть когда выход задерживается на фиксированное количество временных интервалов и прогноз выполняется на более поздние моменты времени. Мы показываем, как снижается точность, когда мы пытаемся предсказать данные о трафике в будущих временных интервалах. Кроме того, мы анализируем влияние количества наблюдений, которые может учитывать сеть LSTM, и продолжительности временных интервалов T : это параметры проектирования, которые необходимо оценить, поскольку они определяют размер памяти сети LSTM и сколько информации о трафике необходимо хранить для точного прогнозирования.

На рис. 5 мы показываем результаты прогнозирования мобильного трафика для двух ячеек: поскольку они расположены в двух разных областях, контролируемые eNodeBs представляют два различных профиля трафика с точки зрения профиля и величины трафика. Мы видим, что прогноз точен на всю неделю, несмотря на колебательное поведение трафика. В этом случае прогноз выполняется на один шаг вперед, что означает, что мы используем фиксированное число прошлых значений ($K = 10$) для прогнозирования трафика для следующего временного интервала.

На рис. 6 и 7 мы оцениваем ошибку предсказания относительно прошлых наблюдаемых значений. Важно также учитывать различные значения для длительности временного интервала T , что влияет на расчет агрегированного трафика $S(T)$ из необработанных трасс LTE. Цифры связаны с первым eNodeB (результаты сопоставимы для eNodeB 2). На рис. 7 мы видим, что NRMSE больше для большей продолжительности, и, как и ожидалось, ошибка уменьшается с большим количеством наблюдений.

Рис. 5. Недельный прогноз трафика для разных eNodeB

Рис. 6. График зависимости ошибки предсказания от количества прошлых наблюдаемых значений

Рис. 7. График зависимости точности прогноза от количества наблюдаемых значений

В данной работе мы изучаем эффективность рекуррентных нейронных сетей применительно к прогнозированию мобильного трафика. Выбор использования сети LSTM обусловлен характеристиками набора данных, так как мы используем многомерную информацию о трафике, которая выводится непосредственно из DCI канала управления LTE. Блоки LSTM преуспевают захватывать временную корреляцию движения даже для отдаленных временных интервалов. Применение прогнозирования трафика с использованием необработанных агрегированных данных из физического канала является фундаментальным в критически важных по времени приложениях и позволяет избежать необходимости в дополнительных ресурсах для обработки данных трафика.

СПИСОК ЛИТЕРАТУРЫ

1. *Bui N., Widmer J.* Owl: a reliable online watcher for lte control channel measurements. ACM All Things Cellular, 2016
2. *Cho K., B. van Merriënboer, Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y.* Learning phrase representations using rnn encoder-decoder for statistical machine translation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. P. 1724–1734.
3. *Andrews J.G., Buzzi S., Choi W., Hanly S.V., Lozano A., Soong A.C., Zhang J.C.* What will 5g be? // IEEE Journal on selected areas in communications. 2014. Vol. 32, no. 6. P. 1065–1082.
4. *Trinh H.D., Bui N., Widmer J., Giupponi L., Dini P.* Analysis and modeling of mobile traffic using real traces. PIMRC, 2017.
5. *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural computation. 1997. Vol. 9, no. 8. P. 1735–1780.
6. *Gers F.A., Schmidhuber J., Cummins F.* Learning to forget: Continual prediction with LSTM – 1999.
7. *Kingma D.P., Ba J.* Adam. A method for stochastic optimization: arXiv preprint arXiv: 1412.6980, 2014.