УДК 512.6:004.7

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ПРОГНОЗИРОВАНИЯ ТРАФИКА В ТЕЛЕКОММУНИКАЦИОННЫХ СИСТЕМАХ

К.М. Руккас, Ю.В. Соляник
Харьковский национальный университет им. В.Н. Каразина

К.А. Овчинников
Харьковский национальный университет радиоэлектроники

Олоту Олуватосин Давид
Харьковский национальный университет им. В.Н. Каразина

Abstract *– This paper presents the results of a comparative analysis of different time series forecasting methods applied to telecommunication systems Internet traffic. To test the accuracy of these methods several experiments were held using real data from public Wi-Fi Internet gateway in different time periods. In addition different time scales (1s, 1min, 10 min) were analyzed. The experiment shows a presence of relatively small but significant linear correlation between first lags which justifies linear model application for network traffic prediction. Two different forecast approaches were used: traffic amount prediction and time series sign classification. Obtained results show the efficiency of non-linear regression models (neural networks and logistic regression) for traffic amount prediction and non-linear regression and decision trees (algorithm C4.5) for sign classification. Also was shown that model efficiency strongly depends from data gathering interval. For sign classification depending on forecasting lookahead the difference between non-linear models and decision trees decreases from 3,15% on 1 step to 0,72% on 20 steps. For both cases logistic regression is the most preferable method due to the balance between efficiency and complexity. Obtained results can be used for increasing the efficiency of automatic network managements systems and in traffic modeling tasks.*

Анотація *– У роботі наведені результати порівняльного аналізу ефективності застосування різних методів прогнозування до часових рядів обсягів інтернет-трафіка телекомунікаційних систем. При проведенні експериментів найбільша точність прогнозу була отримана при використанні нелінійних регресійних моделей часових рядів.*

Аннотация *– В работе приведены результаты сравнительного анализа эффективности применения различных методов прогнозирования к временным рядам объемов интернет-трафика телекоммуникационных систем. При проведении экспериментов наибольшая точность прогноза была получена при использовании нелинейных регрессионных моделей временных рядов.*

Введение

Прогнозирование характеристик трафика является одной из важных задач при построении автоматизированных систем управления телекоммуникационными системами (ТКС). Различные виды коммуникационных услуг и различные конфигурации сетей порождают существенно различающиеся виды трафика [1]. Общеизвестно [1], что, как правило, временные ряды характеристик трафика существенно нестаци-

онарны, т.е. их вероятностные характеристики изменяются во времени. Поэтому моделирование трафика в ТКС представляет собой сложную задачу.

Построение моделей нестационарных временных рядов предполагает выполнение трудоемких задач, таких как предварительная обработка данных, выбор и оценка пригодности моделей. Для частичной автоматизации процесса моделирования трафика необходимы методология и инструментарий, учитывающие специфику предметной области и позволяющие, в зависимости от контекста, построить приемлемое решение. Внедрение этого решения потенциально может позволить полностью автоматизировать процесс управления трафиком.

Технология построения прогноза временного ряда обычно предполагает построение модели ряда, а затем расчет необходимого значения с помощью этой модели. Принято ассоциировать методы прогнозирования с моделями временного ряда. В результате исследований в данной области [1-5] были выявлены такие свойства трафика в ТКС как группировка во времени значительных изменений значений («скачков»), фрактальность, хаотичность, нелинейность, циклические колебания, присущие в разной мере рядам его характеристик в зависимости от контекста. Наибольшее количество публикаций в этой области пришлось на период 1990-х – начала 2000-х годов.

Данные, используемые в этих работах, обычно брались из открытых источников, таких как [6], для того чтобы можно было сравнить результаты. Большинство из этих наборов данных были сформированы в середине 90-х годов. За последние 20 лет характер трафика в ТКС, очевидно, существенно изменился. Выбор модели является одним из центральных вопросов при построении прогноза. В связи с тем, что в случае прогнозирования нестационарных временных рядов нельзя говорить о преимуществе какой-то из моделей без экспериментального подтверждения, следует постоянно обновлять результаты исследований.

В данной работе приведены результаты сравнительного анализа эффективности применения различных методов прогнозирования к временным рядам объемов интернет-трафика ТКС. Наблюдения были получены в 2013-2014 гг.

I. Постановка задачи

На практике чаще [1] интерес представляет не точечная или интервальная оценка абсолютного значения характеристики трафика, а оценка вероятности, с которой это значение окажется в определенном сегменте диапазона своих значений.

Были поставлены следующие задачи:

1. Сравнить эффективность различных методов прогнозирования значения уровня ряда на один шаг вперед.

2. Сравнить эффективность различных методов прогнозирования знака приращения значения элементов ряда за некоторое количество шагов («горизонт прогноза»). В этом случае оценивается, будет ли значение элемента «выше» или «ниже» текущего уровня через фиксированный интервал времени.

II. Основные этапы анализа данных трафика

С практической точки зрения основными этапами построения прогноза временных рядов являются:

1. Постановка задачи.
2. Предварительная обработка данных:
   i. Обработка «пропусков» в наблюдениях;
   ii. Удаление «выбросов» (данных с нетипичными значениями), что представляет сложную задачу в случае трафика в ТКС, для которого свойственны резкие и значительные по величине изменения значений);
   iii. Разделение наборов данных на сегменты для оценки параметров модели и для оценки ее пригодности;
   iv. Преобразования данных (логарифмирование, взятие разностей, сглаживание и др.);
   v. Структурирование данных в формат, соответствующий входным параметрам используемых функций программного обеспечения.
3. Построение модели:
   i. Определение вида модели;
   ii. Определение суб-оптимального количества факторов, влияющих на отклик на основе технологий отбора наиболее значимых факторов и методов понижения размерности входных данных;
   iii. Оценка параметров;
   iv. Оценка пригодности модели (проверка адекватности модели, расчет значения критерия эффективности).
4. Испытания (построение прогнозов на «тестовом множестве») с использованием различных моделей и их ансамблей (совместное использование нескольких моделей).
5. Выбор лучшего решения.

III. Используемые наборы данных. Предварительная обработка

В работе использовались данные входящего и исходящего Wi-Fi трафика, полученного в апреле и мае 2013 г. в лаборатории Харьковского национального университета радиоэлектроники (рис. 1, рис. 2). Наблюдения, измеренные с интервалом в одну секунду, представляют собой объем трафика в битах. Затем эта информация была агрегирована во временные ряды с интервалами времени 1 минута, 5 минут, 10 минут. На рис. 1 и 2 по оси абсцисс отложено время наблюдения (в секундах), по оси ординат – объем трафика (в Мегабитах).

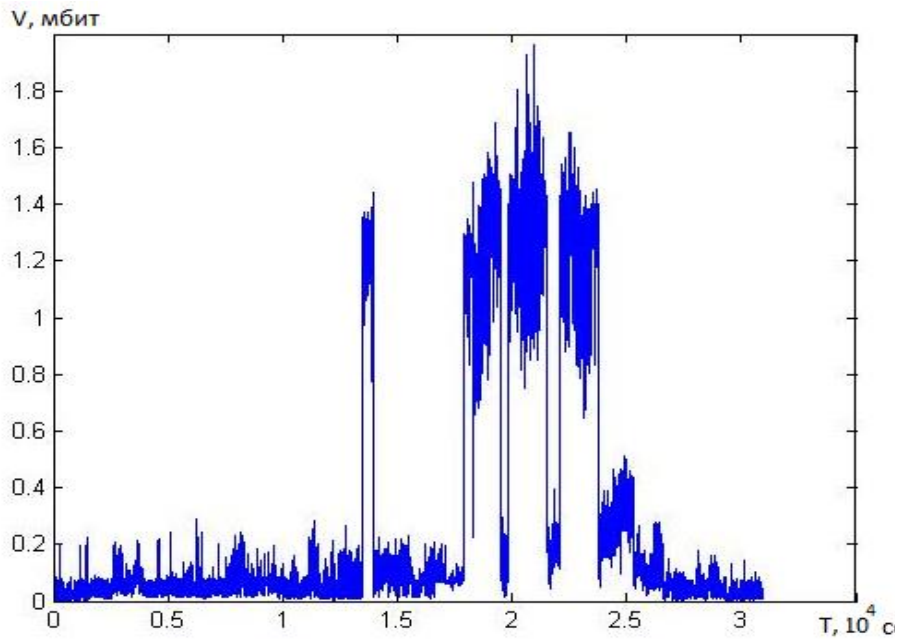

Рис. 1. Временной ряд объемов исходящего трафика с интервалом 1 с. (30.04.2013)

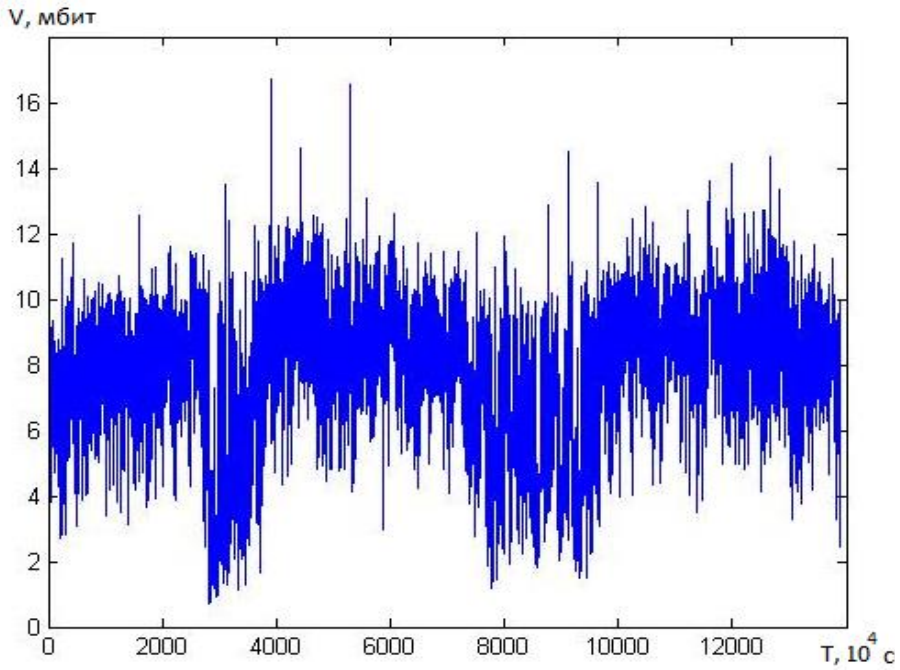

Рис. 2. Временной ряд объемов исходящего трафика с интервалом 1 с. (06.05.2013)

Пропущенные значения наблюдений были заполнены с помощью аппроксимации линейной функцией. Кроме того, были построены таблицы данных, каждая строка состоит из значений факторов (независимых переменных в модели), которые, как предполагается, влияют на отклик, и значения самого отклика (зависимая переменная) – вещественное значение в случае задачи регрессии, и номинальное (категориальное, «метка класса») значение в случае задачи классификации.

Столбцы значений факторов в таблице представляли собой предыдущие (для значения отклика) значения ряда. Их субоптимальное количество подбиралось экспериментально. Эти значения преобразовывались в первые разности (приращения значений за одну единицу времени), для того чтобы привести временной ряд к псевдостационарному виду, более пригодному для построения модели в большинстве случаев [7]. В случае использования адаптивного метода прогнозирования, учитывающего изменение «тенденции» ряда (модель Холта), преобразования не проводились.

Далее значения каждого фактора были нормализованы следующим образом:

$$x_normalized_t = \frac{x_t - \bar{x}}{s},$$

где x_t – значение элемента ряда в момент времени t; x – среднее значение; s – выборочное среднеквадратическое отклонение x (оценка значения квадратного корня из дисперсии) , рассчитываемое по формуле

$$s = \sqrt{\frac{1}{(n-1)}\sum_{t=1}^{n}(x_t - \bar{x})^2},$$

где n – число наблюдений.

Таблица данных в соответствии с традиционной технологией [8] разбивалась на две или три части с сохранением порядка во времени:

• *train/validation* – на сегменте *train* оценивались параметры модели, на множестве *validation* оценивалась ее пригодность в смысле критерия эффективности;

• *train/test/validation* – на сегментах *train*, *test* оценивались параметры модели, принимая во внимание их различие (учитывалась динамика изменения ошибки прогноза) для избегания «переобучения» модели; на сегменте *validation* оценивалась ее пригодность.

Для части экспериментов (результаты которых представлены в табл. 1 и 2) эти сегменты соответствовали наблюдениям различных календарных дней (30.04.2013, 06.05.2013). В остальных случаях (результаты представлены в табл. 3) параметры оценивались на первых 70% данных. Таким образом, эффективность применения определенного метода прогнозирования оценивалась на наблюдениях, которые были получены позже наблюдений, на которых оценивались параметры модели.

IV. Используемые модели и методы прогнозирования

Для решения задач классификации и регрессии в этой работе использовались широко известные модели и методы, входящие в десятку [9] наиболее часто используемых методов *Data Mining*.

Линейная модель используется наиболее часто из-за простоты и относительно небольшой требовательности к ресурсам при оценке параметров, хотя ее применение, как и в случае любой модели, сопряжено с серьезными трудностями при отборе влияющих на отклик факторов:

$$x_t = \tilde{\alpha}_1 x_{t-1} + \ldots + \tilde{\alpha}_k x_{t-k} + \tilde{\alpha}_{k+1} + \tilde{\beta}_1 y_{t-1} + \ldots + \tilde{\beta}_l y_{t-l} + \varepsilon_t; \quad (1)$$

$$\varepsilon_t = N(0,\sigma);$$
$$\sigma = const;$$
$$\mathbf{cov}(x_i \varepsilon_j) = 0, i \neq j,$$

где x_t – элемент временного ряда в момент времени t (наблюдения, эндогенные переменные); y_t – экзогенные переменные в момент времени t (дополнительные «внешние» факторы, влияющие на отклик); $\tilde{\alpha}_i, \tilde{\beta}_i$ – параметры модели; ε_t – случайная ошибка в момент времени t, распределенная по нормальному закону и не имеющая корреляционной связи с наблюдениями в моменты времени $s \neq t$.

Частным случаем (1) (при определенных условиях) является модель авторегрессии и проинтегрированного скользящего среднего (Autoregressive Integrated Moving Average, ARIMA) [7].

Разновидность модели логистической регрессии, используемой в ниже описанных экспериментах, имеет вид

$$y = \frac{1}{1+e^{-z}}, \tag{2}$$

где z – линейная модель (1) временного ряда.

Эта модель часто используется для решения задач классификации. Отображая вещественное число в интервал (0, 1), она позволяет сегментировать отклик с помощью пороговых значений, ставя, таким образом, ему в соответствие метку класса. Также модель (2) часто используется как переходная функция при построении нейронных сетей.

Все чаще при прогнозировании временных рядов используется многослойный персептрон как разновидность нейронной сети прямого распространения. Ее архитектура и возможности широко известны и хорошо описаны [10].

Модель Холта [11] (двухпараметрическое экспоненциальное сглаживание) – адаптивная модель, позволяющая при построении прогноза P ряда X «подстраиваться» к изменениям наклона тренда T и отклонения от уровня S:

$$P_{t+1} = S_t + T_t;$$
$$S_{t+1} = \alpha X_t + (1-\alpha)P_t;$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1};$$
$$0 \le \alpha \le 1;$$
$$0 \le \beta \le 1,$$

где X_t – элемент временного ряда в момент времени t; P_t – прогнозы элементов ряда на момент времени t; α, β – параметры модели Холта.

Методы Naive Bayes [12], Bayes Net [10] (сети Байеса) основаны на байесовском подходе к вероятностному/статистическому анализу.

Алгоритм C4.5 [12] является реализацией одного из самых эффективных методов построения классификаторов типа «деревья решений», использующий информационную метрику (энтропию) как один из критериев выбора локального решения.

Метод AdaBoost [10] позволяет построить ансамбль предикторов, ошибки которых минимально коррелируют. В качестве «базового» предиктора использовались деревья решений.

Следует отметить, что при сравнении эффективности методов прогнозирования не проводился всеобъемлющий анализ адекватности каждой из моделей. В качестве критериев сравнения использовались стандартные расчетные показатели, такие как средняя доля правильно определенных классов (в задачах классификации) и средняя абсолютная ошибка (в задачах прогнозирования уровня ряда). Также отметим, что при построении моделей явно не учитывались некоторые широко известные свойства трафика в ТКС, такие как фрактальность и циклические колебания [1, 2, 5].

V. Результаты эксперимента

Прогноз значения уровня ряда

На рис. 3 и рис. 4 представлены диаграммы автокорреляций и частных автокорреляций [7] первых разностей временного ряда объемов трафика с интервалом 1 с. По оси абсцисс отложено значение автокорреляций различных лагов (autocorr), а по оси ординат – значения лагов (Lag).

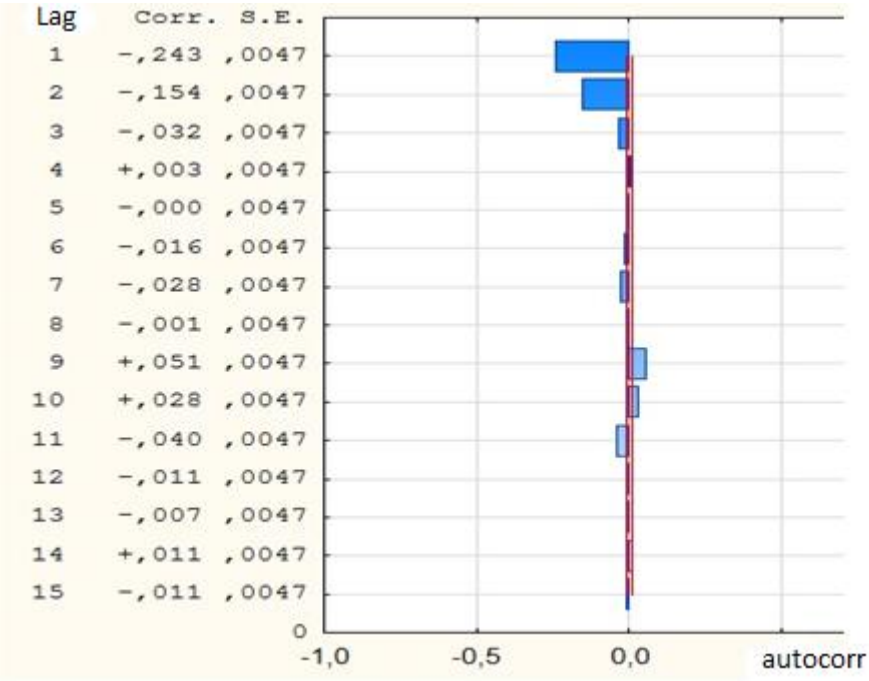

Рис. 3. Коррелограмма первых разностей временного ряда объема трафика с интервалом 1 с

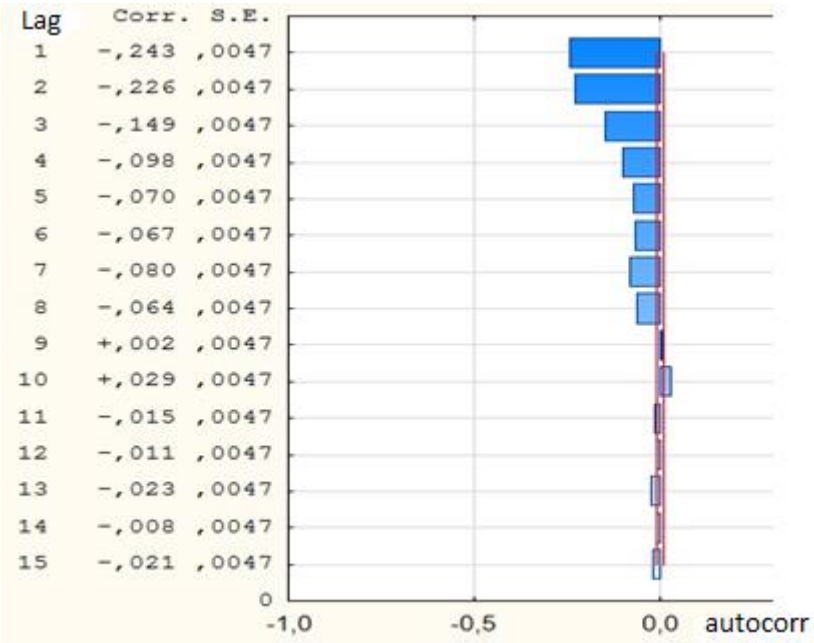

Рис. 4. Диаграмма частных автокорреляций первых разностей временного ряда объема трафика с интервалом 1 с

С помощью этих диаграмм можно убедиться в том, что существует небольшая, но значимая линейная зависимость для нескольких первых лагов, что оправдывает использование линейных моделей временного ряда в данном случае.

В табл. 1 приведены результаты построения прогноза на один шаг вперед для временных рядов объема исходящего трафика с временными интервалами 1 с, 1 мин., 10 мин. В качестве критерия эффективности использовалась средняя абсолютная ошибка (*mean absolute error*, *MAE*), рассчитываемая по формуле

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | \varepsilon_t |,$$

где ε_t – ошибка прогноза на шаге t.

Таблица 1. Средние абсолютные ошибки прогноза на один шаг вперед временного ряда объемов исходящего трафика

Метод /временной интервал	**1 с**	**60 с**	**10 мин.**
Многослойный персептрон	0,0203	0,0199	0,0123
Модель Холта	0,0230	0,0141	0,0192
Линейная модель	0,0208	0,0455	0,0571
Среднее абсолютных значений первых разности ряда	0,3596	0,3027	0,3213

В последней строке табл. 1 приведены средние абсолютных значений первых разностей (приращений) ряда, для того чтобы можно было оценить, какова была бы ошибка, если бы в качестве прогноза использовалось предыдущее значение. Как уже

было указано выше, данные были нормализованы. Из приведенной в табл. 1 информации можно сделать вывод, что с увеличением интервала ряда точность линейной модели ухудшается, и что наименьшую ошибку прогноза обеспечивают многослойный персептрон и модель Холта.

Рассмотрев рис. 5, можно объяснить понижение качества линейной модели с ростом интервала временного ряда (значимые автокорреляции для интервала ряда 60 с отсутствуют или значительно ниже, чем для интервала 1 с). Сравнение результатов двух других моделей позволяет говорить об их разной чувствительности к «зашумленности» рядов.

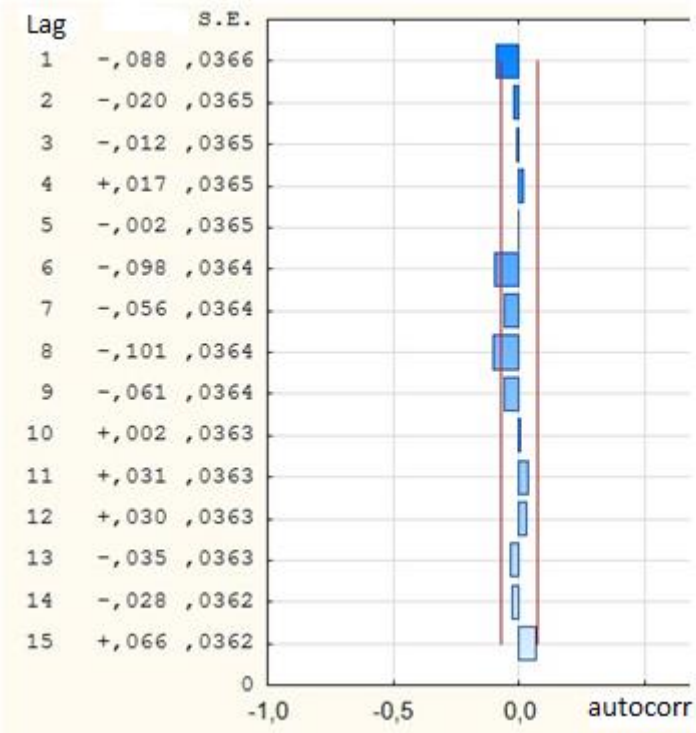

Рис. 5. Коррелограмма первых разностей временного ряда объема трафика с интервалом 60 с

Прогноз знака приращения уровня ряда

В табл. 2 приведены результаты решения задачи классификации знака приращения уровня ряда с горизонтом прогноза один шаг. Использовался входящий трафик с интервалом наблюдения 1 с. В качестве критерия эффективности использовалась доля верно определенных классов (в процентах), рассчитываемая по формуле $100 \cdot \frac{m}{n}$, где n – количество попыток определить класс; m – количество попыток, при которых класс был определен верно.

По данным табл. 2 можно сделать вывод, что наилучший прогноз дают нелинейные модели (логистическая регрессия, многослойный персептрон).

Таблица 2. Результаты прогноза на один шаг вперед знака приращения уровня ряда входящего трафика с интервалом наблюдения 1 с

Метод	**процент верно определенных классов**
Логистическая регрессия	62,52
Многослойный персептрон	61,81
AdaBoost	56,93
C4.5	56,52
Bayes Net	52,10
Naïve Bayes	49,73

В табл. 3 приведены результаты классификации знака приращения уровня ряда с различными горизонтами прогноза. Использовался входящий трафик с интервалом наблюдения 5 мин.

Таблица 3. Результаты прогноза знака приращения уровня ряда с интервалом наблюдения 5 мин. Расчетная величина – процент верно определенных классов

Метод / Горизонт прогноза	**1**	**5**	**10**	**20**
Многослойный персептрон	61,15	72,20	73,31	81,66
Логистическая регрессия	59,43	68,53	72,18	81,81
C4.5	56,28	64,92	73,91	81,09
Bayes Net	54,73	58,04	60,74	71,64
AdaBoost	53,52	58,63	62,82	77,23
Naïve Bayes	52,93	56,80	56,28	55,65

В этом случае с ростом горизонта прогноза все модели повышают точность. Можно предположить, что это связано с высоким «уровнем шума» для рядов с меньшим временным интервалом. Наиболее высокую точность прогнозирования показали нелинейные модели (многослойный персептрон и модель логистической регрессии).

Выводы

В статье был проведен численный эксперимент, целью которого было сравнить эффективность различных методов прогнозирования в применении к временным рядам характеристик трафика в ТКС. Результаты эксперимента позволяют говорить о том, что для использованных наборов данных практически во всех случаях наибольшая точность прогноза была получена при использовании нелинейных регрессионных моделей временных рядов (нейронные сети, логистическая регрессия), что согласуется с исследованиями прошлых лет [1-5].

При прогнозировании значения уровня ряда (табл. 1) эффективность применения нейронной сети значительно превышает эффективность применения линейной модели, и это различие чувствительно к изменению временного интервала временного ряда. Так, для интервала наблюдения 1 с. оно составляет

((0,0208-0,0203)/0,0203)*100%=2,46%,

а для интервала 10 с. –

((0,0571-0,0123)/0,0123)*100%=364,22%.

При сравнении результатов применения адаптивной модели Холта и линейной модели соответствующие числа равны 9,57% (результаты применения модели Холта хуже для интервала 1 с.) и 197,34% (результаты применения модели Холта лучше для интервала 10 с.).

При решении задачи классификации (прогнозировании знака приращения уровня ряда за период) нелинейные регрессионные модели также показывают лучшие результаты, однако различие между эффективностью методов зависит от величины горизонта прогноза. Так, из результатов, приведенных в табл. 3, видно, что, при изменении значения горизонта прогноза от 1 до 20 разница в эффективности применения логистической регрессии и деревьев решений (алгоритм C4.5) снижается от 59,43%-56,28%=3,15% до 81,.81% -81,09=0,72%.

Отметим, что эффективность применения логистической регрессии в среднем ненамного уступает, а иногда и превышает эффективность применения нейронной сети. В связи с тем, что времени на оценку параметров модели логистической регрессии в среднем требуется значительно меньше, это может быть решающим фактором при выборе эффективного и приемлемого в смысле требований к вычислительным ресурсам решения.

Практическая применимость результатов должна быть рассмотрена в соответствии с критериями эффективности работы конкретной системы управления сетью. Дальнейшая работа может быть направлена на создание автоматизированных систем моделирования трафика в ТКС с целью понижения трудоемкости и повышения эффективности управления компьютерными сетями.

Список литературы:

1. *Шелухин О.И., Тенякшев А.М., Осин А.В.* Фрактальные процессы в телекоммуникациях: Монография / Под ред. О.И. Шелухина. – М.: Радиотехника, 2003. – 480 с.

2. *Leland W.E., Taqqu M.S., Willinger W., Wilson D.V.* On the self-similar nature of ethernet traffic // IEEE/ACM Transactions of Networking. – 1994. – Vol. 2, No.1. – P. 1-15.

3. *Chabaa S., Zeroual A., Antari J.* Identification and Prediction of Internet traffic Using Artificial Neural Networks // Intelligent Learning Systems & Applications. – 2010. – No.2. – P. 147-155.

4. *Gowrishankar S., Satyanarayana P. S.* A Time Series Modeling and Prediction of Wireless Network Traffic // International Journal of Interactive Mobile Technologies (iJIM). – 2009. – Vol.4, No.1. – P. 53-62.

5. *Петров В.В., Платов В.В.* Исследование самоподобной структуры телетрафика беспроводной сети // Радиотехнические тетради, №3. – М.: ОКБ МЭИ, 2004. – С. 58-62.

6. The Internet Traffic Archive [Электронный ресурс]. – Режим доступа: http://ita.ee.lbl.gov.

7. *Box, G.E.P., Jenkins G.M., Reinsel G.C.* Time Series Analysis: Forecasting and Control, 4th edition, Prentice Hall, 2008. – 810 p.

8. *Webb A.R.* Statistical Pattern Recognition, Second Edition, John Wiley & Sons, Ltd., 2002. – 495 p.

9. *Wu X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G.J., Ng A., Liu B., Yu P.S., Zhou Z.H., Steinbach M., Hand D.J., Steinberg D.* Top 10 Algorithms in Data Mining // Knowledge and Information Systems. – 2008. – Vol. 14, No. 1. – P. 1-37.

10. *Рассел С., Норвиг П.* Искусственный интеллект: современный подход, 2 изд. – М.: Издательский дом «Вильямс», 2007 . – 1408 с.

11. *Лукашин Ю.Л.* Адаптивные методы краткосрочного прогнозирования: Учеб. пособие. – М.: Финансы и статистика, 2003. – 416 с.

12. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007. – 384 с.