پیش بینی ترافیک آینده با استفاده از مدل های پنهان مارکوف

چکیده

تخمین و پیش بینی حجم ترافیک شبکه موضوع تحقیقاتی مهمی است که توجه مداوم انجمن شبکه و انجمن یادگیری ماشین را جلب کرده است. اگر چه کار زیادی بر روی تخمین و پیش بینی ماتریس ترافیک با استفاده از مدل های سری زمانی، تجزیه ماتریس با رنک پایین وجود دارد، بر اساس آنچه ما و همکاران می دانیم کارهای کمی برای بررسی این مسئله که آیا ما می توانیم حجم شبکه مبتنی بر برخی آمارگان (آماره های) های ترافیک که جمع آوری آنها، کم هزینه تر هستند مانند جریان شمارش flow count ، تخمین زده یا پیش بینی کنیم. در این مقاله، مدلی برای ارتباط بین حجم ترافیک و آمارگان (آماره های) های ساده مانند جریان flows با استفاده مدل پنهان مارکف پیشنهاد می دهیم که بر اساس آن می توانیم از اندازه گیری مستقیم حجم داده اجتناب کنیم اما در عوض حجم ترافیک پنهان شده مبتنی بر آن آمارگان (آماره های) ساده جریان flow که به وسیله برخی تکنیک‌های طراحی جمع آوری شده اند، تخمین زده و پیش بینی کنیم. سادگی و تاثیرگذاری روش پیشنهادی را با استفاده از تعدادی شبه شبیه سازی و نتایج تجربی حاصل از داده واقعی نشان می دهیم.

1. مقدمه

تخمین و پیش بینی حجم ترافیک شبکه[1] یک مسئله مهم تحقیقاتی شبکه است. تخمین و پیش بینی دقیق حجم ترافیک ، به ویژه ماتریس ترافیک ، برای کنترل مسیریابی شبکه ، کنترل ازدحام ، تخصیص منابع شبکه و برنامه ریزی بلند مدت سودمند است و به این ترتیب در انجمن شبکه و انجمن یادگیری ماشین توجه زیادی را به خود جلب کرده است. در کارهای موجود عمدتا دو جریان اصلی از تحقیق وجود دارد. جریان اصلی اول فرض می کند

[1] Network traffic volume

که در هر بازه زمانی مشخص ، حجم ترافیک مجموع[۲] (تعداد بایت) بین یک جفت منبع مقصد معین که می تواند اندازه گیری شود، اتفاق می افتد و سپس یک مدل آنالیز سری زمانی مانند مدل های خطی شامل AR، ARMA، ARIMA، [6] FARIMA، [7]، [16]، [14] و مدل های غیر خطی شامل ANN، RNN، [10] GARCH، [17]، [2]، [8] برای پیش بینی ترافیک آینده استفاده شده است. محدودیت این دسته از رویکردها این است که لازم است به طور مستقیم حجم ترافیک بازه های زمانی قبلی را به منظور پیش بینی حجم های ترافیک برای بازه زمانی آینده اندازه گیری نماییم. با این وجود، اندازه گیری مستقیم حجم داده برای عملی شدن به ویژه در شبکه با سرعت خیلی بالا بسیار هزینه بر است، و بنابراین اگرچه این رویکرد ساده است اما در عمل، مقیاس‌پذیر نیست . مسیر اصلی دیگر رویکردها معمولاً توموگرافی شبکه نامیده می شوند [3] ، [1] ، [9] ، [4] که مکمل اولین رویکرد مسیر اصلی است. ایده توموگرافی شبکه برای تخمین حجم ترافیک شبکه مبتنی بر مشاهدات دیگری مانند استفاده از لینک link utilization است. استفاده از لینک link utilization[۳]، حجم ترافیک مجموع از جریان هایی flow است که از طریق آن لینک عبور می کنند. در نتیجه، معمولاً یک سیستم خطی معین (قطعی) برای توصیف رابطه بین کاربرد لینک و حجم ترافیک پنهان وجود دارد. با این حال ، یکی از محدودیت های مهلک رویکرد توموگرافی شبکه این است که سیستم خطی همیشه نامعین است زیرا در یک شبکه تعداد لینک ها به مراتب کمتر از تعداد جفت های منبع مقصد است. بازیابی حجم ترافیک پنهان با استفاده ازمقدار محدودی از استفاده در لینک link utilizations بسیار دشوار است.

در این مقاله ، با توجه به محدودیت های قوی در کارهای موجود، احتمال حدس زدن حجم ترافیک را بر اساس برخی آمارگان (آماره های) های جریان مانند تعداد جریان های بازه زمانی معین که جمع آوری بسیار راحت تری دارند را بررسی می کنیم. می دانیم، کار ما در استخراج وابستگی بین شمارش های جریان flow count و حجم جریان به منظور تخمین و پیش بینی حجم ترافیک، پیشگام است. پیشنهاد می کنیم از مدل پنهان مارکف برای تشریح ارتباط شمارش جریان flow count و حجم جریان و هم چنین رفتار پویای موقت هر دو استفاده کنیم.

---

[2] aggregated
[3] link utilization

ما از الکوریتم های بسیار جدیدی مانند قانون کرنل بیز و شبکه های عصبی بازگشتی با واحد حافظه طولانی کوتاه مدت (واحد LSTM) برای آموزش مدل و استفاده از مدل برای پیش بینی ترافیک آینده استفاده می کنیم.

در ادامه این مقاله ، در بخش II، درباره کارهای موجودی که برای تخمین و پیش بینی ترافیک شبکه نقش داشته اند، بحث می نماییم؛ در بخش III، مدل مارکوف پنهان خود را برای تخمین و پیش بینی ترافیک آینده با استفاده از قانون Kernel Bayes[5] [15] و همچنین شبکه عصبی بازگشتی پیشنهاد می کنیم. در بخش IV، آزمایش هایی را با استفاده از داده نیمه شبیه سازی شده و داده واقعی ترافیک شبکه انجام می دهیم؛ در بخش V از این مقاله نتیجه گیری می نماییم.

2. تخمین و پیش بینی ترافیک شبکه

در دهه های گذشته، کارهای زیادی برای حل مشکل تخمین و پیش بینی ترافیک شبکه منتشر شده است. همانطور که در مقدمه بحث شد، آن کارها عمدتاً به دو دسته اصلی تقسیم می شوند. در یک مورد فرض می کنیم که ما می توانیم ترافیک شبکه مجموع در فواصل زمانی متوالی را مشاهده کنیم و یک مدل ریاضی برای پیش بینی ترافیک آینده را با روشی ساده به وجود آوریم[6]، [7]، [16]، [14]، [10] ، [17] ، [2] ، [8]. دسته دیگر روشهایی که به آن توموگرافی شبکه گفته می شود از اندازه گیری مستقیم ترافیک شبکه بین هر دو موردی که مورد مذکور جلوگیری می کند، اما در عوض سعی می کند با استفاده از برخی از link utilizations، حجم ترافیک پنهان را بازیابی کند. در این بخش، به طور مختصر در مورد فرمول ها و همچنین محدودیت های این دو گروه روش ها صحبت خواهیم کرد.

A. پیش بینی غلتان با استفاده از مشاهدات قبلی

این گروه از روشها [3] ، [1] ، [9] ، [4] فرض می کنند که ما قادریم حجم ترافیک را به صورت دنباله مشاهده کنیم. هدف ما پیش بینی ترافیک آینده بر اساس مشاهدات قبلی است. اساس این دسته از روشها، خود شباهتی در ترافیک شبکه است. به طور کلی، می توان از فرمول زیر برای توصیف روند پیش بینی استفاده کرد:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \cdots, \mathbf{x}_{t-p+1}, \boldsymbol{\epsilon}_t, \cdots, \boldsymbol{\epsilon}_{t-q+1}) + \boldsymbol{\epsilon}_{t+1}, \qquad (1)$$

x_t حجم ترافیک منتقل شده در زمان t از آن دسته از جفت های مقصد است که مورد توجه است، ϵ_t خطای پیش بینی در زمان t است ، p تعداد مشاهدات قبلی است که برای پیش بینی استفاده می شود و q تعداد خطاهای پیش بینی قبلی است که برای تصحیح پیش بینی استفاده می شوند.

مرحله آموزش برای یادگیری بهترین عملکردی است که به حداقل می رساندخطای پیش بینی به شرح زیر است:

$$f^* = \underset{f}{argmin}\, \mathbb{E}[(\mathbf{x}_{t+1} - f(\mathbf{x}_t, \cdots, \mathbf{x}_{t-p+1}, \boldsymbol{\epsilon}_t, \cdots, \boldsymbol{\epsilon}_{t-q+1}))^2].$$

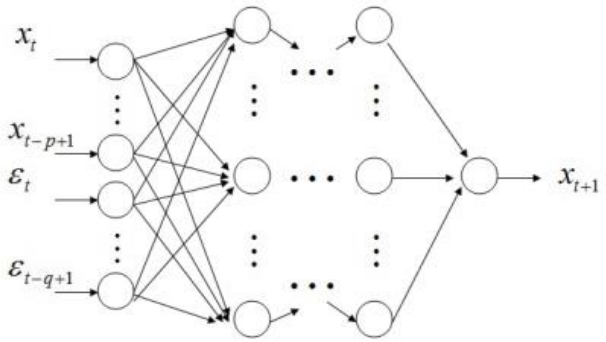
مدل های زیادی برای تقریب f استفاده می شوند. یک مورد ساده مدل خطی مانند ARMA و انواع آن مانند ARIMA و FARIMA است. در مدل های ARMA ، رابطه بین پیش بینی کننده و متغیر هدف به سادگی با استفاده از یک مدل خطی به شرح زیر توصیف می شود:

$$\mathbf{x}_{t+1} = \sum_{i=0}^{p-1} \alpha_i \mathbf{x}_{t-i} + \sum_{j=0}^{q-1} \beta_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_{t+1}, \qquad (2)$$

که α_i و β_j ضرایبی هستند که به راحتی می توان رگرسیون حداقل مربعات یاد گرفته شوند.

مدلهای خطی به راحتی قابل پیاده سازی هستند و تفسیر خوبی دارند و بنابراین در بسیاری از مسائل تحلیل سری زمانی در کارهای واقعی به طور گسترده ای مورد استفاده قرار می گیرند. با این حال، نشان داده شده است مدل های خطی برای توصیف برخی از رفتارهای غیرخطی از ترافیک شبکه کافی نیستند. برای انعطاف پذیری بیشترمدل، همچنین کارهایی با استفاده از شبکه عصبی مصنوعی (ANN) برای تقریب عملکرد غیرخطی تابع f وجود دارد.

ANN تقریب زن بسیار قوی تابع غیر خطی با تعداد کافی داده شده از نورون های پنهان است. همانطور که در شکل 1 نشان داده شده است، ما مشاهدات پیشین و خطاهای پیش بینی را به شبکه عصبی و خروجی های شبکه عصبی حجم ترافیک پیش بینی شده آینده نمایش می دهیک . مرحله آموزش شبکه عصبی تنظیم وزنهای اتصالات بین دو لایه مجاور نورون ها به منظور به حداقل رساندن خطاهای پیش بینی است. بازگشت به عقب با استفاده از کاهش گرادیان دسته ای و کاهش گرادیان تصادفی معمولاً برای آموزش یک شبکه عصبی استفاده می شود.



شکل 1- شبکه عصبی مصنوعی

اگرچه ایده پیش بینی غلتان با استفاده از مشاهدات قبلی، ساده و کارآمد به نظر می رسد ، اما محدودیت اصلی این است که باید در فواصل زمانی متوالی حجم ترافیک را جمع آوری کند که می تواند به خصوص در یک شبکه با سرعت بالا در مقیاس بزرگ بسیار گران باشد. برای جلوگیری از اندازه گیری مستقیم حجم ترافیک، روش هایی وجود دارد که به آن توموگرافی شبکه گفته می شود که برای تخمین حجم ترافیک از داده های link utilization پیشنهاد شده است و سپس از ترافیک تخمین زده شده برای انجام پیش بینی همانطور که در قسمت بعدی توضیح داده شده استفاده می کند.

B. توموگرافی شبکه

ایده توموگرافی شبکه بهره برداری از رابطه بین داده بار لینک و تقاضای ترافیک در بین هاست های انتهایی شبکه است. با استفاده از X_t یک بردار جمع آوری شده از تمام حجم های ترافیکی منتقل شده در شکاف زمانی t بین هر دو هاست انتهایی شبکه و یک ماتریس روتینگ (مسیریابی) توسط A که حاوی اطلاعات مسیریابی است، یعنی $A_{i,j} = 1$ بدین معنی است که لینک i به مسیری تعلق دارد که جفت مقصد j برای انتقال ترافیک خود استفاده می کند. در غیر اینصورت $A_{i,j} = 0$. ما همچنین توسط t Y_t یک بردار را نشان می دهیم که تمام بارهای لینک را در شکاف زمانی t جمع آوری می کند. سپس می توانیم رابطه بین بار لینک[۴] (لینک لود) و حجم ترافیک را با سیستم خطی زیر فرمول بندی کنیم:

[4] Link load

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t, \ \forall t. \qquad (3)$$

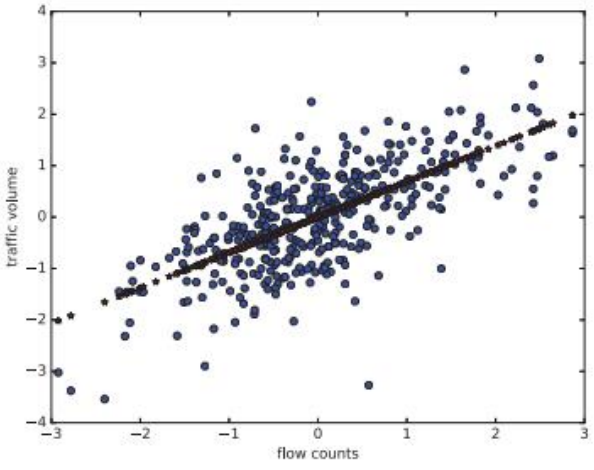
توجه داشته باشید که فرض می کنیم در طول دوره مشاهده ، ماتریس روتینگ تغییر داده نشده است. حتی با این فرض، سیستم حل توصیف شده در معادله 3 بسیار مشکل است زیرا سیستم نامعین است زیرا تعداد لینک ها معمولا به مراتب کمتر از تعداد جفت های هاست های انتهایی در یک شبکه هستند.

تحقیقات گسترده ای با استفاده ازسنجش فشرده، الگوریتم های امید ریاضی- بشینه سازی برای حل این سیستم وجود دارد. به طور کلی ، آن الگوریتم ها پیچیده هستند و نتایج رضایت بخش نیست.

با درک محدودیت های قوی در کارهای موجود ذکر شده، چارچوبی جدید برای برآورد حجم ترافیک بر اساس برخی از آماگان های ساده سطح جریان را با استفاده از مدل های پنهان مارکوف در بخش زیر پیشنهاد می کنیم.

3. حدس زدن و پیش بینی حجم ترافیک با استفاده از HMM ها

در این بخش، امکان حدس زدن[5] و پیش بینی حجم ترافیک را بر اساس برخی از آمارگان (آماره های) ساده سطح جریان ساده مورد بحث قرار می دهیم. این ایده مبتنی بر مشاهداتی است که وابستگی آماری قوی بین آن آمارگان (آماره های) ساده سطح جریان و کل حجم ترافیک وجود دارد که توسط شکل زیر که با تجزیه و تحلیل سری زمانی از ترافیک شبکه واقعی بدست آمده است، نشان داده شده است.



شکل 2: رابطه بین تعداد جریان و حجم جریان

[5] inferring

شکل 2 رابطه بین تعداد جریان در شکاف های زمانی مختلف و حجم کل جریان مربوطه را (بر حسب بایت) نشان می دهد. در اینجا سری های زمانی را مانند آنهایی که دارای میانگین و انحراف معیار[6]های صفر دارند را نرمالیزه می کنیم. ما می توانیم ببینیم که می توان از این ارتباط برای حدس زدن حجم بر اساس شمارش جریان (flow count) استفاده کرد زیرا همبستگی[7] مهمی بین شمارش جریان (flow count) و حجم ترافیک وجود دارد.

A. آمارگان (آماره های) ساده جریان و تکنیک طراحی

ما استدلال می کنیم که جمع آوری آمارگان (آماره های) ساده جریان بسیار ارزان تر از اندازه گیری مستقیم حجم ترافیک است. در اینجا می توان آمارگان جریان را به صورت آمارگان زیر تعریف کنیم:

$C_{f.t}$ تعداد جریان (flow) در فاصله زمانی t

$C_{tcp.t}$ تعداد جریان های TCP در فاصله زمانی t

$C_{R(i).t}$ تعداد جریان ها با استفاده از تعداد پورت موجود در بازه R(i)، $\forall i$ در زمان t

علاوه بر مواردی که در اینجا تعریف می کنیم، هیچ محدودیتی در استفاده از اطلاعات بیشتری که برای تخمین حجم ترافیک مفید است، وجود ندارد. ما تمام آمارگان (آماره های) جریان را که به راحتی جمع آوری می شود ، در یک بردار مشاهدات $y_t = [C_{f.t}, C_{tcp.t}, C_{R(i).t}, \dots]^T$ قرار می دهیم.

جمع آوری این آمارگان (آماره های) سطح جریان، به طور طبیعی، مسئله ای آیتم های مجزای قابل شمارش یک رشته داده است که به طور گسترده در مقالات مورد مطالعه قرار گرفته است.

B. مدل های پنهان مارکف

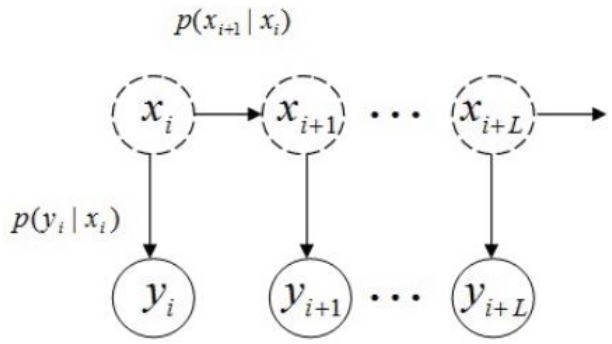
مدل های پنهان مارکوف معمولا از مدل های فضا-حالت[8] تغییز ناپذیر با زمان[9] به شرح زیر استفاده می کنند:

$$p(\mathbf{X}, \mathbf{Y}) = \pi(\mathbf{x}_0) \prod_{i=0}^{T} p(\mathbf{y}_i|\mathbf{x}_i) \prod_{i=0}^{T-1} p(\mathbf{x}_{i+1}|\mathbf{x}_i), \qquad (4)$$

---

[6] standard deviations
[7] correlation
[8] state-space
[9] time invariant

که x_iمتغیر پنهان و y_i متغیر مشاهده شده است ، $p(x_i + 1 \mid y_i)$ احتمال انتقال[10] است که رفتار پویای سیستم را توصیف می کند و $p(y_i \mid x_i)$ احتمال انتشار[11] است که توصیف می کند سیستم چگونه مشاهدات را بر اساس متغیرهای پنهان تولید می کند. در مسئله ما ، حجم ترافیک متغیر پنهان x_iاست و آمارگان (آماره های) جریان متغیر مشاهده شده y_i است و بنابراین $p(x_i + 1 \mid x_i)$ چگونگی تغییرحجم ترافیک در طول زمان و $p(y_i \mid x_i)$ رابطه بین حجم ترافیک و آمارگان جریان مانند تعداد شمارش جریان (flow count) را توصیف می کند.



شکل 3- مدل پنهان مارکف

به طور کلی ، $p(x_i + 1 \mid x_i)$ و $p(y_i \mid x_i)$ نامعلوم هستند و بنابراین لازم است آن ها را با تقریب زدن به وسیله برخی رویکردهای پارامتری یا یادگیری آن از داده ها تخمین بزنیم. در این مقاله فرض می کنیم که ما قادر به جمع آوری برخی داده های آموزشی $(x_0.y_0.x_1.y_1 \ldots . x_2.y_2)$ هستیم به گونه ای که قادر به یادگیری احتمال انتقال و احتمال انتشار هستیم. در مسئله شبکه سازی، فرض می کنیم که با توجه به مقدار محدود budget ذخیره سازی و قابلیت اندازه گیری، می توانیم برخی ازpacket traces را جمع آوری کنیم که می تواند به ما چگونگی نموهای ترافیک شبکه و ارتباط بین حجم و آمارگان جریان مانند شمارش جریان (flow count) یاد دهد. این امر یک بار انجام می‌شود و برای تخمین و پیش بینی آینده، ما فقط نیاز داریم آمارگان جریان را جمع آوری نماییم که بر اساس آن حجم ترافیک را خدس زده و پیش بینی کنیم.

فرض کنید ما یک دنباله از آمارگان (آماره های) جریان مشاهده شده جدید $(\tilde{y}_0.\tilde{y}_1 \ldots . \tilde{y}_t)$ را داریم، می خواهیم متغیر پنهان متناظر $\tilde{x}_t$ را حدس بزنیم، یعنی:

---

[10] transition probability

[11] emission probability

$$\tilde{\mathbf{x}}_t \sim p(\mathbf{x}_t|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0), \qquad (5)$$

علاوه بر این، ما می خواهیم حجم ترافیک آینده را پیش بینی کنیم:

$$\tilde{\mathbf{x}}_{t+1} \sim p(\mathbf{x}_{t+1}|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0). \qquad (6)$$

در اینجا ما قانون Kernel Bayes را دنبال می کنیم تا یک تخمین نقطه ای از $\tilde{x}_t$ و $\tilde{x}_{t+1}$ بدست آوریم. ایده اصلی تعبیه کردن احتمال انتقال و احتمال انتشار به RKHS به عنوان تعبیه میانگین شرطی[12] به شرح زیر است:

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t) \mapsto \hat{C}_{X_{+1}X}(\hat{C}_{XX} + \epsilon I)^{-1}, \qquad (7)$$

$$p(\mathbf{y}_t|\mathbf{x}_t) \mapsto \hat{C}_{YX}(\hat{C}_{XX} + \epsilon I)^{-1}, \qquad (8)$$

که :

$$\hat{C}_{UV} = \frac{1}{L}\sum_{i=1}^{L} k(\cdot, U_i) \otimes k(\cdot, V_i),$$

که $k(.,.)$ یک تابع کرنل مانند تابع نمایی درجه دوم $k(x.\acute{x}) = \exp\left(-(x-\acute{x})/\sigma^2\right)$ است. فرض کنید $p(x_t \mid \tilde{y}_t \ldots . \tilde{y}_0)$ را به فضای کرنل هیلبرت بازتولید شونده[13] [11] به صورت تعبیه میانگین شرطی $\hat{m}_{x_t|\,\tilde{y}_t \ldots . \tilde{y}_0}$ تعبیه کردیم [12]، سپس می توانیم $\tilde{x}_t$ را با پیدا کردن مقداری که شرط زیر را مینیمم می کند، تخمین بزنیم:

$$\tilde{\mathbf{x}}_t = \underset{\mathbf{x}}{argmin}\, \|k(\cdot, \mathbf{x}) - \hat{m}_{\mathbf{x}_t|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0}\|^2_{\mathcal{H}}.$$

علاوه بر این می توانیم $p(x_{t+1} \mid \tilde{y}_t \ldots . \tilde{y}_0)$ را به فضای کرنل هیلبرت باز تولیده شونده به صورت تعبیه میانگین شرطی $\hat{m}_{x_{t+1}|\,\tilde{y}_t \ldots . \tilde{y}_0}$ با استفاده از قانون Kernel Bayes به صورت زیر تعبیه کنیم

$$\hat{m}_{\mathbf{x}_{t+1}|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0} = \hat{C}_{X_{+1}X}(\hat{C}_{XX} + \epsilon I)^{-1}\hat{m}_{\mathbf{x}_t|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0},$$

زیرا:

$$p(\mathbf{x}_{t+1}|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0) = \int p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0)d\mathbf{x}_t$$

---

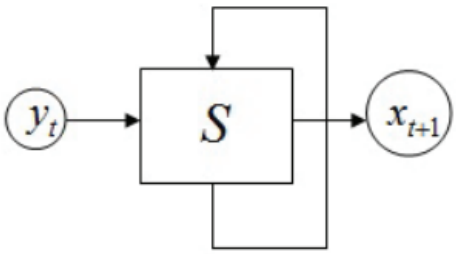
[12] condition mean embeddings

[13] Reproducing Kernel Hilbert Space

و در RKHS، انتگرال به آسانی ضرب ماتریسی را کاهش می دهد. به طور مشابه، می توانیم با پیدا کردن مقداری که شرط زیر را به حداقل می رساند $\tilde{x}_{t+1}$ را پیش بینی کنیم:

$$\tilde{\mathbf{x}}_{t+1} = \underset{\mathbf{x}}{argmin} \, \|k(\cdot, \mathbf{x}) - \hat{m}_{\mathbf{x}_{t+1}|\tilde{\mathbf{y}}_t, \cdots, \tilde{\mathbf{y}}_0}\|^2_{\mathcal{H}}.$$

پیچیدگی محاسباتی KBB، $O(n^3)$است که n در اندازه نمونه آموزش است. پیچیدگی محاسباتی می‌تواند به $O(nr^2)$ نیز کاهش یابد که r << n کاردینالیتی زیرمجموعه ی رگرسیون ها ست. برای اطلاعات بیشتر، لطفا به [5] ، [15] مراجعه کنید.



شکل 4- پیش‌بینی ترافیک آینده با استفاده از مدل RNN

یک جایگزین استفاده از شبکه های عصبی بازگشتی به منظور پیش بینی حجم ترافیک آینده بر اساس شمارش جریان (flow counts) است. بر خلاف شبکه های عصبی پیشخور که در آن هیچ ارتباطی بین نورون ها در همان لایه پنهان وجود ندارد، RNN ها نورون ها را در همان لایه های پنهان ارتباط دارند و این بدان معنی است که RNN ها حافظه داخلی را برای ذخیره حالت های قبلی نورون ها حفظ می کنند و علاوه بر ورودی های لایه قبلی، حالت های قبلی آن نورون ها را به خودشان بازخورد می‌کنند چنان که برای پردازش دنباله دلخواه ورودی ها بسیار مناسب است و بنابراین برای کارهایی مانند تحلیل سری های زمانی کاربردی است.

در مسئله ما ، ما فقط نیاز به آموزش RNN داریم که در هر بازه زمانی t ، شمارش جریان (flow count) را در نظر می گیرد و سپس تخمینی از حجم ترافیک آینده را در فاصله زمانی t + 1 تولید می کند. هدف از فاز آموزش، یادگیری وزن های بهینه اتصالات در RNN است به گونه ای که خطای پیش بینی به حداقل برسد. در صورت لزوم می توان از شبکه عصبی بازگشتی عمیق[۱۴] استفاده کرد.

[14] Deep recurrent neural network

4. آزمایشات

در این بخش، آزمایش هایی با استفاده از داده های نیمه شبیه سازی و داده های ترافیک شبکه واقعی انجام می دهیم تا امکان حدس زدن و پیش بینی میزان ترافیک شبکه را بر اساس آمارگان (آماره های) ساده جریان مانند شمارش جریان (flow count) نشان دهیم. در آزمایش های زیر، هر دو سری زمانی را به گونه ای که دارای میانگین صفر و انحراف معیار صفر داشته باشند، نرمالیزه می کنیم.

A. شبه شبیه سازی

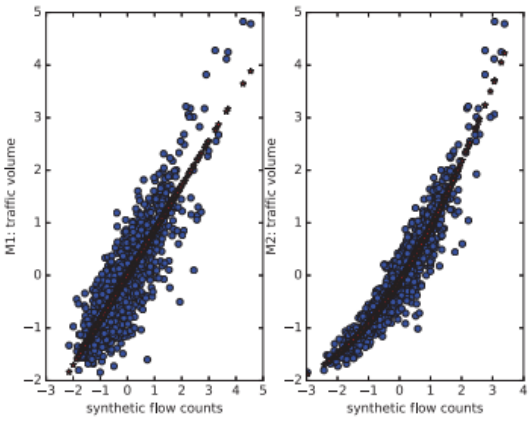
در این بخش، نیمه شبیه سازی را انجام می دهیم که از داده های بنچمارک (الگو) عمومی به نام داده های 2004 Abilene از Internet استفاده می کنیم.

این مجموعه داده شامل میانگین های 24 هفته 5 دقیقه ای برای 12 روتر (ماتریس 12 * 12) است. در این آزمایش، ما فقط از ترافیک درایه های (3 ، 3) ماتریس استفاده می کنیم.

x_tحجم ترافیک را در فاصله زمانی t و y_t شمارش جریان (flow count) متناظر را نشان می‌دهد. شمارش های جریان (flow counts) ازحجم ترافیک را با استفاده از مکانیزم های زیر تولید می‌کنیم (توزیع های شرطی) که انواع مشخصی از غیرخطی بودن و تصادفی بودن را نشان می دهد.

- M1: $y_t = 0.01 * (x_t + 0.05 * \xi_1 + 0.05 * \xi_2)$,
- M2: $y_t = 0.01 * ((x_t)^{0.1} + 0.25 * \xi_1 + 0.25 * \xi_2)$,

که ξ_1 از توزیع استاندارد گاما و ξ_2 از توزیع استاندارد گاوسی پیروی می کند.



شکل 5- تعداد جریان مصنوعی برحسب حجم ترافیک

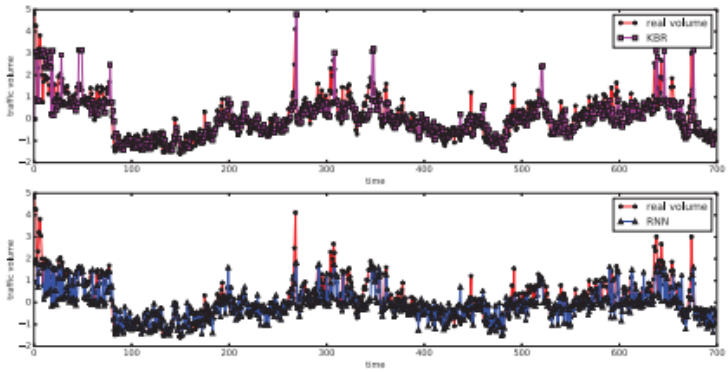
آزمایشات را تحت تنظیمات مختلف آموزش و اندازه نمونه های آزمایشی (S_{tr},S_{tst}) = (200,700) (300,600),(400, 500), انجام می دهیم. نتایج در جدول های زیر برای M1 و M2 خلاصه شده است.

جدول I و II خطای میانگین مربعات[15] پیش بینی را در هر دو سناریو برای توزیع شرطی شمارش جریان (flow count) با توجه به حجم جریان که تحت تنظیمات مختلف در مورد اندازه نمونه آموزش و اندازه‌گیری نمونه آزمایش است را نشان می دهد.

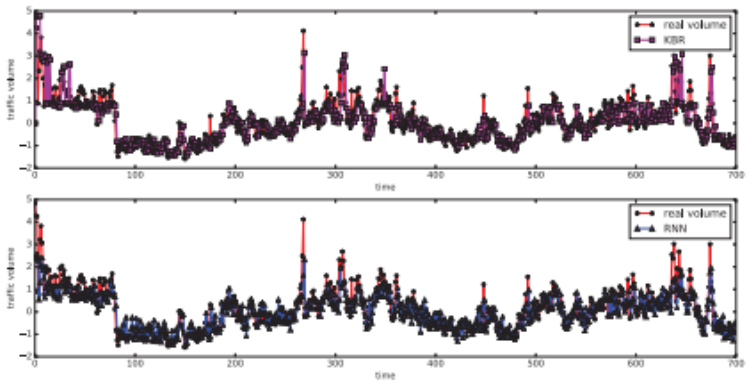
جدول 1- خطای میانگین مربعات پیش‌بینی برای M_1

Algo	(200, 700)	(300, 600)	(400, 500)
KBR	0.2215	0.1684	0.1991
RNN	0.2396	0.2151	0.2538

درمی یابیم که خطای پیش بینی بسیار ناچیز است زیرا در اینجا سری زمانی را نرمالیزه می کنیم که سری زمانی اصلی با واریانس واحد است.



شکل 6 – پیش‌بینی حجم ترافیک بر اساس تعداد جریان تحت M_1



شکل 7 – پیش‌بینی حجم ترافیک بر اساس تعداد جریان تحت M_2

[15] Mean Square Error

شکل 6 و 7 حجم ترافیک پیش بینی شده بر حسب حجم ترافیک واقعی برای هر دو شرط M1 با نمونه آموزش برابر با 200 نشان می دهد. می توانیم ببینیم حجم ترافیک پیش بینی شده بسیار نزدیک به مقدار واقعی است.

B. ترافیک شبکه واقعی

در این بخش ، آزمایش هایی را با استفاده از ترافیک واقعی شبکه اینترنت انجام می دهیم. ما کل trace را به 400 فاصله زمانی بر اساس تایم استمپ[۱۶] های رکورد های جریان و شمارش تعداد جریان ها و هم چنین تعداد کل بایت ها در طول هر فاصله زمانی تقسیم می کنیم.

رابطه بین flow countها و کل حجم ترافیک در شکل 2 نشان داده شده است. ما همچنین آزمایش‌هایی را با استفاده از اندازه نمونه آموزش مختلف و اندازه نمونه آزمایش (S_{tr},S_{tst}) = (100,300) (200، 200) و (100، 300) انجام دادیم.

جدول 2- میانگین مربعات خطای پیش بینی برای M2

Algo	(200, 700)	(300, 600)	(400, 500)
KBR	0.1624	0.1454	0.1599
RNN	0.1686	0.1545	0.1608

جدول 3- خطای میانگین مربعات پیش‌بینی

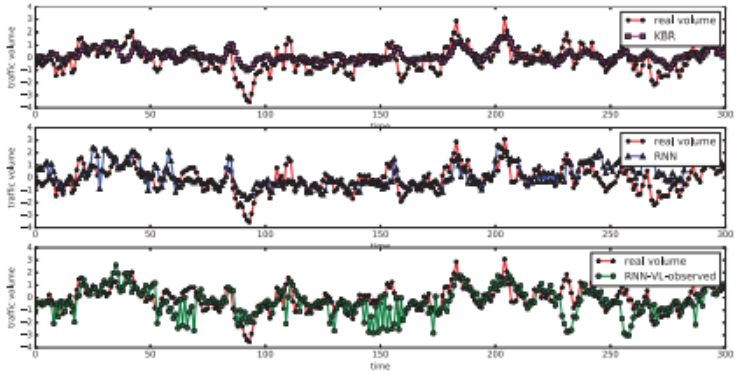
Algo	(100, 300)	(200, 200)	(300, 100)
KBR	0.3545	0.3092	0.2926
RNN	0.4285	0.4126	0.3441
RNN (volume observed)	0.6524	0.2202	0.1517

در این آزمایش، همچنین دقت پیش بینی مبتنی بر شمارش جریان (flow count) را با دقت پیش بینی مبتنی بر حجم ترافیک مشاهده شده در زمان های قبلی مقایسه می کنیم. توجه داشته باشید که در این مقاله، از ایده پیش بینی مبتنی بر آماره های ساده مانند شمارش های جریان (flow counts) حمایت می‌کنیم زیرا اندازه‌گیری مستقیم حجم ترافیک بسیار گران است. در این آزمایش ، ما پیش بینی‌های سری زمانی را مبتنی بر مشاهدات پیشین حجم ترافیک، فقط به منظور نشان دادن فاصله بین پیش بینی مبتنی بر مشاهدات نویزی و پیش بینی

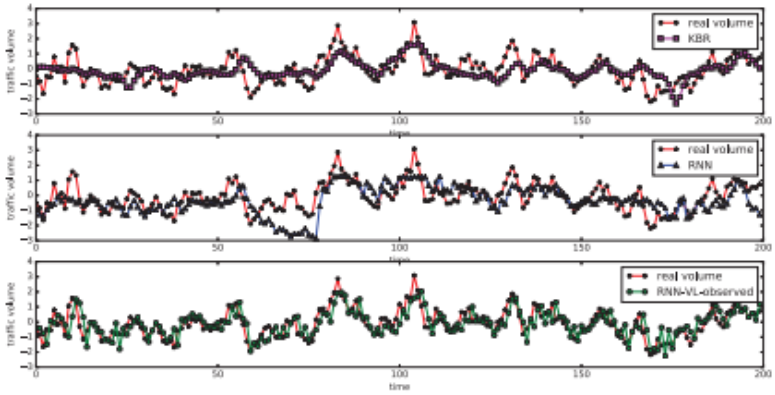
[16] timestamp

مبتنی بر اطلاعات کامل (بدون نویز) ارائه می دهیم. ما دوباره از RNN استفاده می کنیم اما ورودی RNN حجم ترافیک مشاهده شده پیشین غیراز تعداد جریان (flow count) پیشین است.

دقت های پیش بینی در جدول III نشان داده شده است.

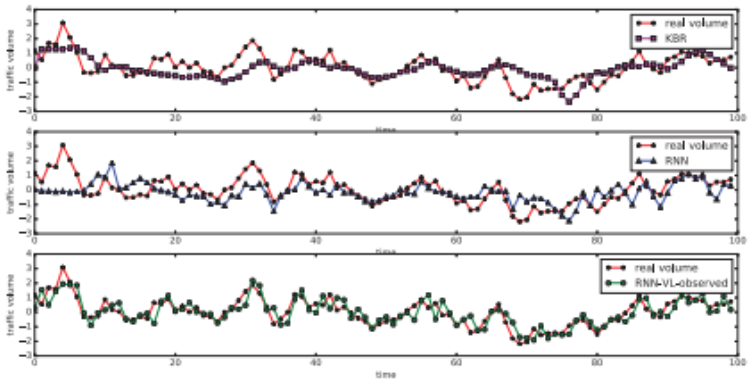


شکل 8: پیش بینی حجم ترافیک براساس تعداد Flow (1:3 = آزمایش: آموزش)



شکل 9: پیش بینی حجم ترافیک براساس تعداد Flow (2:2 = آزمایش: آموزش)

به طور کلی می توان دید ، وقتی اندازه نمونه بزرگتر است، خطای پیش بینی کوچکتر می شود. حتی هنگامی که نسبت بین اندازه نمونه آزمایشی با اندازه نمونه آموزی 3 باشد ، خطاهای پیش بینی برای KBR و RNN کمتر از 0.5 است که کمتر از نصف واریانس سری های زمانی هستند. می توان مشاهده مرد بین پیش بینی های نویزی و پیش بینی مبتنی بر اطلاعات کامل (بدون نویز)، زیاد نیست که نشان می دهد باید راه حلی وجود داشته باشد که دقت پیش بینی و هزینه های مانیتورینگ را متعادل کند.



شکل 10 - شکل 8: پیش بینی حجم ترافیک براساس تعداد Flow (3:1 = آزمایش: آموزش)

شکل 8 ، 9 و 10 نیز سری های زمانی پیش بینی شده برحسب سری زمانی واقعی تحت تنظیمات مختلفی از نسبت های اندازه نمونه آزمایشی و آموزشی نشان می دهد. که به طور کلی می توان مشاهده کرد سری های زمانی پیش بینی شده با حالت واقعی کاملا مطابقت دارد.

5. نتیجه گیری

در این کار، چگونگی استفاده از چندین تکنیک یادگیری ماشین از جمله مدل پنهان مارکوف مبتنی بر فانون Kernel Bayes و همچنین شبکه عصبی بازشگتی را برای تخمین حجم ترافیک آینده و همچنین پیش بینی حجم ترافیک آینده بر اساس برخی از آماره های سطح جریان ساده که می تواند با روش راحت تر یعنی با استفاده از تکنیک های طراحی sketch جمع آوری می شود، توضیح دادیم. این رویکرد از اندازه گیری مستقیم حجم ترافیک جلوگیری می کند و بنابراین از لحاظ پیچیدگی و نیازمندی ذخیره سازی[17] بسیار کم هزینه تر است. این امر به ویژه در شبکه های بسیار پرسرعت (مقیاس بزرگ) مفید است که در آن اندازه گیری مستقیم حجم ترافیک برای همه جفت های مقصد مبدا تقریبا غیرممکن است و تخمین حجم ترافیک شبکه از بار لینک (لود لینک) بسیار دشوار است. انجام نیمه شبیه سازی و آزمایشات با استفاده از داده های ترافیک شبکه واقعی، نشان می دهد که استفاده از آمارگان سطح جریان ساده مانند شمارش جریان، اطلاعات مفیدی را برای پیش بینی حجم ترافیک

[17] storage requirement

فراهم می‌کند. در کار بعدی، ما قصد داریم چارچوب پیشنهادی را برای مانیتورینگ شبکه واقعی و مهندسی ترافیک استفاده کنیم.

موارد باقی مانده دیگری وجود دارد که باید به آن اشاره کرد. یکی از آنها این است که آیا وابستگی بین حجم ترافیک و آمارگان سطح جریان ساده مانند شمارش جریان در کلیه شبکه ها مانند WAN و ترافیک مرکز interdata به اندازه کافی قابل توجه است. مسئله دوم عدم ثبات در ترافیک شبکه است. از آنجا که ترافیک شبکه به صورت پویا در حال تغییر است که بدان معنی است که رفتارهای تابع انتقال و تابع انتشار نیز می تواند تغییر کند. در این حالت، لازم است الگوریتم های یادگیری آنلاین را برای KBR و همچنین RNN را توسعه دهیم به گونه ای که مدل خود را با ترافیک پویای شبکه تنظیم و سازگار کند. سؤال قابل بحث سوم این است که علاوه برشمارش جریان چه آمارگان دیگری از سطح جریان می تواند برای بهبود دقت پیش بینی ممکن است، استفاده شود.

REFERENCES

[1] J. Cao, D. Davis, S. Vander Wiel, and B. Yu. Time-varying network tomography: router link data. Journal of the American statistical association, 95(452):1063–1075, 2000.

[2] S. Chabaa, A. Zeroual, J. Antari, et al. Identification and prediction of internet traffic using artificial neural networks. Journal of Intelligent Learning Systems and Applications, 2(03):147, 2010.

[3] A. Chen, J. Cao, and T. Bu. Network tomography: Identifiability and fourier domain estimation. IEEE Transactions on Signal Processing, 58(12):6029–6039, 2010.

[4] M. H. Firooz and S. Roy. Network tomography via compressed sensing. In Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, pages 1–5. IEEE, 2010.

[5] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule. In Advances in neural information processing systems, pages 1737–1745, 2011.

[6] N. K. Hoong, P. K. Hoong, I. K. Tan, N. Muthuvelu, and L. C. Seng. Impact of utilizing forecasted network traffic for data transfers. In Advanced Communication Technology (ICACT), 2011 13th International Conference on, pages 1199–1204. IEEE, 2011.

[7] P. K. Hoong, I. K. Tan, and C. Y. Keong. Bittorrent network traffic forecasting with arma. arXiv preprint arXiv:1208.1896, 2012.

[8] W. Junsong, W. Jiukun, Z. Maohua, and W. Junjie. Prediction of internet traffic based on elman neural network. In 2009 Chinese Control and Decision Conference, pages 1248–1252. IEEE, 2009.

[9] G. Liang and B. Yu. Maximum pseudo likelihood estimation in network tomography. IEEE Transactions on Signal Processing, 51(8):2043–2053, 2003.

[10] D.-C. Park and D.-M. Woo. Prediction of network traffic using dynamic bilinear recurrent neural network. In 2009 Fifth International Conference on Natural Computation, volume 2, pages 419–423. IEEE, 2009.

[11] B. Scholkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.

[12] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. IEEE Signal Processing Magazine, 30(4):98–111, 2013.

[13] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 961–968. ACM, 2009.

[14] A. R. Syed et al. Forecasting network traffic load usingwavelet filters and seasonal autoregressivemoving average model. International Journal of Computer and Electrical Engineering, 2(6):979, 2010.

[15] Y. Xu. Kernel bayes rule. Journal of Machine Learning Research, 14, 2013.

[16] S. Yantai, Y. Minfang, Y. Oliver, L. Jiakun, and F. Huifang. Wireless traffic modeling and prediction using seasonal arima models. IEICE transactions on communications, 88(10):3992–3999, 2005.

[17] E. Yu and C. R. Chen. Traffic prediction using neural networks. In Global Telecommunications Conference, 1993, including a Communications Theory Mini-Conference. Technical Program Conference Record, IEEE in Houston. GLOBECOM'93., IEEE, pages 991–995. IEEE, 1993.