بخش سه

همان‌طور که در بخش دو مورد بحث قرار گرفت، روش طبقه بندی ترافیک موجود از مشکل کاربرد روز صفر به دلیل عدم وجود نمونه‌های ترافیکی روز صفر در بخش تعلیم دسته بند رنج می برد. چگونگی رسیدن به یک مدل ترافیک روز اول کافی تبدیل به یک سؤال کلیدی برای حل پایه‌ای این مسأله شده است. کار ما توسط رصدهایی که داده‌های شبکه ای شامل ترافیک روز صفرم را بدون برچسب می‌کرد تهییج شد. هدف ما ساخت یک دسته بند سریع با خارج کردن نمونه‌های روز صفر و وارد کردن آن‌ها در مرحله تعلیم است.

این بخش یک شمای طبقه بندی ترافیکی سریع را برای درگبر شدن با کاربردهای روز صفر نمایش می دهد. شکل ۱ یک چارچوب جدید RTC را نشان می دهد. ماژول مهم در چارچوب پیشنهادی وجود دارند: تشخیص ناشناس،دسته بندی ترافیکی بر پایه ٬کیسه جریان ها٬ (BOF) و به روزرسانی های سیستم. ماژول شناخت ناشناس با هدف تشخیص اتوماتیک نمونه‌های جدید ترافیک روز صفر در یک مجموعه از ترافیک بدون برچسب جمع شده به صورت تصادفی از شبکه هدف است. ماژول دسته بندی ترافیکی بر اساس (BOF) نمونه‌های از پیش برچسب خورده ی تعلیمی و نمونه‌های ترافیک روز صفر را به عنوان ورودی برای ساخت یک دسته بند برای دسته بندی ترافیکی سریع می گیرد. برای دستیابی به دسته بندی ریز دانه، ماژول یک به روز رسانی سیستمی می‌تواند به صورت هوشمندانه ترافیک روز صفر را تحلیل کند و کلاس‌های جدیدی را برای تکمیل دانش سیستم بسازد. در مقاله حاضر ما یک کاربرد از RTC را ارایه می‌دهیم که الگوریتم جنگل تصادفی و k-means برای انجام طبقه بندی نظارت شده و آموزش بدون نظارت به کار برده می‌شوند (خوشه بندی).

a. کشف ناشناس

ما یک روش دو مرحله‌ای کشف ناشناس را برای استخراج نمونه‌های ترافیکی روز صفر از یک مجموعه از ترافیک شبکه بدون برچسب که برای شمای RTC حیاتی هستند پیشنهاد می دهیم. روش دو مرحله‌ای در الگوریتم ۱ خلاصه شده است. مرحله اول تشخیص بر اساس k-means خوشه های ترافیکی روز صفر است. دومین مرحله استخراج نمونه‌های روز صفر با استفاده از جنگل تصادفی است.

با داشتن مجموعه های تعلیمی از پیش برچسب خورده ی {v1,….,vn} و یک مجموعه ی بدون برچسب Tu، ما تقریباً تعدادی از مدل های روز صفر را از Tu با استفاده از یک ایده ی نیمه نظارتی برای اولین مرحله خارج می کنیم. مدل های برچسب دار و بدون برچسب برای تغذیه ی الگوریتم خوشه بندی با هم ادغام می شوند. خوشه بندی k-means با هدف بخش بندی جریان ترافیک به k خوشه به گونه ای که (k < |T|)، C={C1,…..,Ck} تا مجموع مربعات درون خوشه ای را به حداقل برساند. الگوریتم سنتی k-means یک تکنیک پالایش تکرار شونده را به کار می گیرد. با داشتن یک مجموعه اولیه ازk مرکز (centroids,) انتخاب شده ی تصادفی، الگوریتم با تناوب بین مرحله ی واگذاری و مرحله ی به روز رسانی ادامه می یابد(40). در مرحله ی واگذاری هر جریان به خوشه با نزدیکترین میانگین واگذار می شود.



در مرحله ی به روز رسانی، میانگین های جدید برای تبدیل شدن به مراکز جریان های درون خوشه محاسبه می شوند. با انتخاب یک k بزرگ (25)،(41) ما خوشه های ترافیکی با خلوص بالا {C1,…..Ck} را به دست می آوریم. نمونه های تعلیمی از پیش برچسب خورده سپس می توانند برای تعیین خوشه های ترافیکی روز صفر مورد استفاده قرار گیرند. قاعده به این صورت است

* اگر یک خوشه شامل هیچ نمونه ی از پیش برچسب خورده ای نباشد، یک خوشه ی نرافیکی روز صفر است.

هر چند به عبارت ساده یک k بزرگ باعث یک نرخ TP بالا و همچنین یک نرخ FP بالا از تشخیص ناشناس می شود که خلوص نمونه های ناشناس تشخیص داده شده را جدا تحت تاثیر قرار می دهد.

در مرحله دوم ما ساخت یک دسته بند جنگل تصادفی را برای رسیدگی به این موضوع پیشنهاد می دهیم. یک کلاس ناشناس عمومی برای نشان دادن ترکیب کاربردهای روز صفر پیشنهاد می شود. مجموعه ی نمونه های روز صفر V گرفته شده در مرحله اول به صورت موقت به عنوان مجموعه ی تعلیمی برای کلاس ناشناس عمومی مورد استفاده قرار می گیرد. بنابراین ما یک مسئله ی خاص طبقه بندی چند کلاس شامل N کلاس شناخته شده و یک کلاس ناشناس داریم. سپس مجموعه های از پیش برچسب خورده {V1,….,Vn} و مجموعه نمونه های روز صفر موقت V برای تعلیم یک دسته بند جنگل تصادفی fc1 ترکیب می شوند. جنگل تصادفی با توانایی خوب تعمیم بخشی، کارایی طبقه بندی عالی ای را در کار پیشین در طبقه بندی ترافیک نشان داده است. ما در ادامه fc1 را برای دسته بندی جریان ها در Tu برای دست یابی به یک مجموعه با خلوص بالا از نمونه های روز صفر U به کار بردیم. به صورت خاص برای تضمین خلوص نمونه های روز صفر ما یک روش دسته بندی جدید را که هم بستگی جریان (4) را در ترافیک دنیای واقعی در نظر می گیرد به کار گرفتیم. این مورد با جزییات در بخش III-B توضیح داده خواهد شد.

B. دسته بندی ترافیک بر اساس BoF

برای دسته بندی سریع ترافیک ما بیشتر پیشنهاد یک روش دسته بندی جدید که هم بستگی جریان را در ترافیک شبکه جهان واقعی در نظر می گیرد و جریان های همبسته را به جای دسته بندی تکی با همدیگر دسته بندی می کند.

الگوریتم 2 روش پیشنهاد شده ی دسته بندی ترافیک بر اساس BoF را نشان می دهد. با داشتن مجموعه های تعلیمی از پیش برچسب خورده {V1,….,Vn} و مجموعه نمونه های روز صفر U ساخته شده با ماژول کشف ناشناخته، ما می توانیم دسته بند fc2 را برای دسته بندی کلاس-(N+1) بسازیم. fc2 قادر است که ترافیک روز صفر را به یک کلاس کلی ناشناس دسته بندی کند. در ادامه کار قبلیمان (4) ما هم بستگی جریان را در فرایند دسته بندی ترافیک برای بهتر کردن دقت تشخیص وارد می کنیم. هم بستگی جریان می تواند با یک الگوریتم ابتکاری 3-تاپل کشف شود (30)،(42)،(43). یعنی در یک دوره کوتاه از زمان جریان هایی که یک IP مقصد، پورت مقصد و پروتکل انتقال یکسان را باهم به اشتراک می گذارند با کاربرد/پروتکل یکسان ساخته می شوند. برای راحتی دسته بندی ترافیک ما از "کیسه جریان ها"( bag of flows) برای مدلسازی هم بستگی جریان استفاده کردیم. یک BoF می تواند با X={x1,….,xg} به گونه ای که xi نماینده ی iامین جریان در BoF است تعریف شود. دسته بندی یک BoF می تواند با جمع کردن پیش بینی جریان تولید شده توسط یک دسته بند معمولی شناخته شود. در مقاله حاضر دسته بند جمع شده ی fbof(X) می تواند به این صورت بیان شود:



به گونه ای که fc2 نشان دهنده ی دسته بند جنگل تصادفی و Θ روش رای اکثریت است (44). برای BoF X ما g پیشبینی جریان داریم yx1,…..yxg که توسط fc2 تولید شده است (g تعداد جریان ها در X است). پیشبینی های جریان می توانند به سادگی تبدیل به رای شوند



سپس قاعده انتخاب مرکب به این صورت است:



در نتیجه تمام جریان ها در X به w1 دسته بندی می شوند. دسته بندی ترافیک بر اساس BoF همچنین برای کشف ناشناس در بخش III-A به کار می رود.

در اینجا ما توجیه فورمال (رسمی) منافع هم بستگی جریان برای دسته بندی ترافیک را ارایه می دهیم. در کار قبلی (4) ما یافتیم که دقت دسته بندی ترافیک بر اساس آمار جریان (flow-statistics-based) می تواند به صورت قابل توجهی با ترکیب چندین جریان هم بسته بهتر شود. برای مطالعه نظری و با داشتن آن BoF، X={x1,….,xg} ما یک پیشبینی مرکب با استفاده از متوسط پیشبینی های ایجاد شده در هر جریان ساختیم. بر اساس نظریه تصمیم گیری بیز، قاعده ترکیب متوسط می تواند به قاعده رای اکثریت تحت چارچوب نظری کیتلر تبدیل شود.

اگر ما یک مسئله دسته بندی را در نظر بگیریم، به گونه ای که ما برای پیشبینی یک احتمال پسینی (*posteriori*) تلاش می کنیم و فرض می کنیم که f یک مدل پیشبینی کننده ی تعلیم دیده است، پیشبینی مرکب به این صورت داده می شود:



فرض می کنیم که تایع صحیح احتمال پسینی که سعی می کنیم پیشبینی کنیم توسط p(Xm) داده شده است، خروجی هر کدام از جریان های تصادفی می توانند به عنوان ارزش صحیح به علاوه خطا به فرم زیر نوشته شود



مجموع مربعات متوسط خطا می تواند به صورت زیر نوشته شود



به گونه ای که Xm توزیع خود را دارد و E[.] گویای انتظار با توجه به توزیع آن است. خطای متوسط ایجاد شده توسط هر جریان به صورت زیر است



به صورت مشابه خطای مورد انتظار برای BoF به صورت زیر است



فرض می کنیم که متوسط خطا برابر صفر است و غیر هم بسته است یعنی E[e(Xm)] = 0 و E[e(Xm)e(Xl)] = 0 و m /= l در این صورت ما به دست می آوریم



این نتیجه پیشنهاد می دهد که خطای پیشبینی جریان می تواند با یک ضریب M با استفاده از یک مدل ساده ی بر اساس BoF کاهش یابد.

مطالعه ی بیشتر در مورد Ebof در بخش VI با کم کردن فرض مستقل ارایه خواهد شد.

C. به روز رسانی سیستم

با کشف ناشناس و دسته بندی ترافیک بر اساس BoF شمای پیشنهاد شده ترافیک روز صفر را در حال اجرای دسته بندی ترافیک تشخیص می دهد. ماژول به روز رسانی سیستم برای به دست آوردن دسته بندی ترافیک روز صفر با ریز دانه پیشنهاد شده است. هدف، یادگیری کلاس های جدید در ترافیک روز صفر تشخیص داده شده و تکمیل دانش سیستم است. توانایی یادگیری کلاس های جدید شمای پیشنهاد شده را نسبت به روش دسته بندی ترافیک معمولی متفاوت می کند.

فرایند یادگیری کلاس های جدید در الگوریتم 3 نمایش داده شده است. با داشتن یک مجموعه از ترافیک روز صفر،Z که خروجی دسته بندی ترافیک بر اساس BoF است، ما خوشه بندی k-means را برای به دست آوردن خوشه های {C1,…..Ck} انجام می دهیم. برای هر خوشه ما به صورت تصادفی چندین نمونه جریان (مثلا سه) را برای بازرسی دوطرفه انتخاب می کنیم. برای تضمین خلوص بالای مجموعه های تعلیمی جدید، استراتژی اجماع برای تولید یک پیشبینی اقتباس می شود. اگر تمام جریان های انتخاب شده نشان دهنده ی یک کاربرد/پروتکل جدید باشند ما یک کلاس جدید می سازیم و از جریان های درون خوشه به عنوان داده های تعلیمی آن استفاده می کنیم. برای کلاس جدیدی که در طول به روز رسانی سیستم ایجاد شده است، جریان های درون خوشه به مجموعه تعلیمی آن کلاس افزوده می شود. هنگامی که بازرسی خوشه کامل شد، کلاس های تازه پیدا شده درون مجموعه کلاس های شناخته شده اضافه می شود و مجموعه داده ی تعلیمی بر اساس آن بزرگتر می شود. این بدین معنی است که سیستم دسته بندی توانایی یادگیری کلاسهای جدید را دارد. سیستم به روز رسانی شده می تواند در کاربرد های بیشتری مورد استفاده قرار گیرد و دسته بندی ریزدانه تری به دست آید.

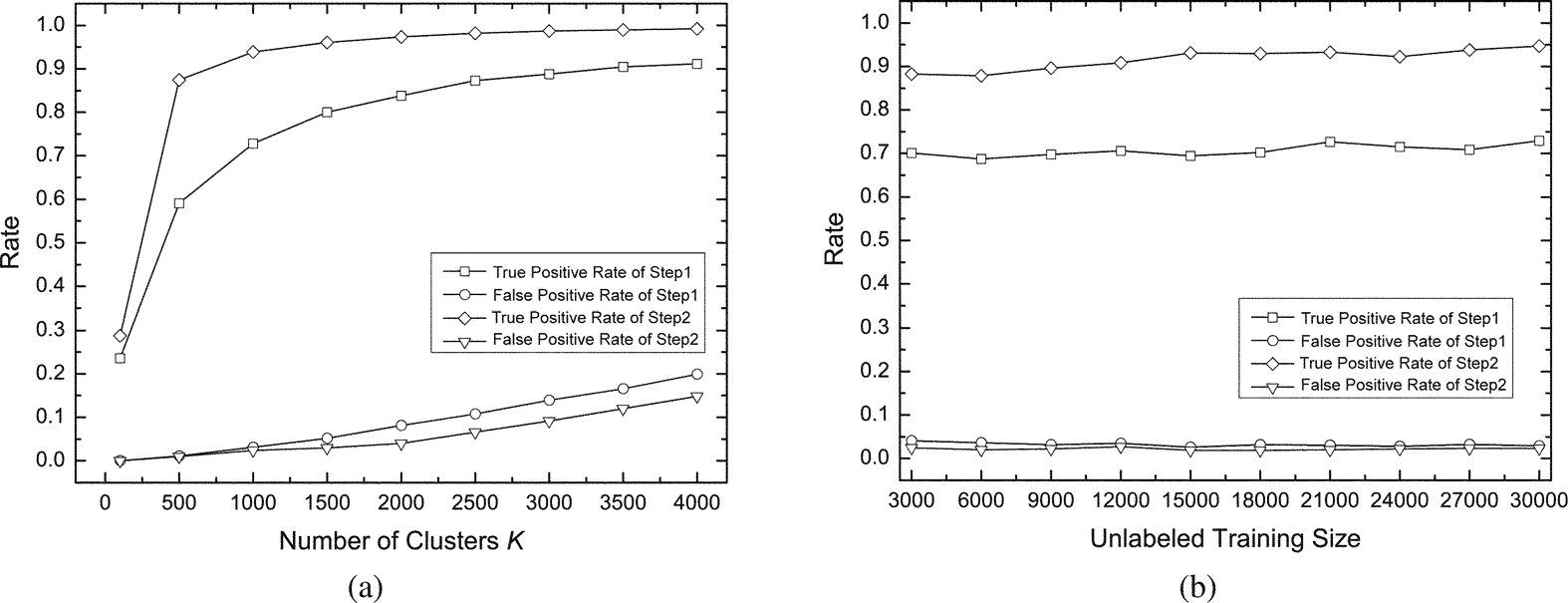
به روز رسانی های مکرر بر اساس تحقیق قبلی (35) ضروری نیست. اگر ترافیک روز صفر دسته بندی شده هر تغییر قابل توجهی را در کاربردها تشخیص دهد، به روزرسانی سیستم برای حفظ دسته بند RTC اجرا می شود. بحثهایی در مورد حفظ دسته بند در بخش VI-B ارایه شده است.

در فرایند معرفی شده در بالا نمونه های تعلیمی برای کلاس های جدید می توانند شامل نویز باشند زیرا خوشه های ترافیک 100% خالص نیستند. این مشکل ممکن است دقت دسته بندی کلاسهای شناخته شده را تحت تاثیر قرار دهد. برای حل این مشکل ما کاربرد یک استراتژی دو مرحله ای دسته بندی را پیشنهاد می دهیم.

در مرحله اول دسته بند کلاسهای-(N+1) به دست آمده قبل از به روز رسانی سیستم می تواند برای انجام دسته بندی ترافیک مورد استفاده قرار گیرد. به صورت ایده آل ترافیک روز صفر به یک کلاس ناشناس عمومی دسته بندی می شود. در دومین مرحله نمونه های تعلیمی برای کلاسهای جدید به دست آمده در طول یک به روز رسانی سیستم می تواند برای تعلیم یک دسته بند جدید استفاده شود و این دسته بند، ترافیک را در کلاس ناشناس عمومی به درون کلاس های جدید ریز دانه دسته بندی می کند. مزیت استراتژی دو مرحله ای دسته بندی این است که کارایی کلاسهای شناخته شده تحت تاثیر قرار نمی گیرد. در این معنا سرعت سیستم دسته بندی ترافیک می تواند بهتر شود.

IV. بهینه سازی پارامتر

تنظیمات یک پارامتر یک چالش قابل توجه برای یک روش دسته بندی ترافیک است که تکنیک یادگیری ماشین را به کار می برد. ما مشاهده کردیم کارایی شمای RTC پیشنهاد شده بستگی به تاثیر کشف ناشناس دارد. در کشف ناشناس دو پارامتر وجود دارند: k که تعداد خوشه های ایجاد شده توسط k-means را مشخص می کندو Tu که اندازه یک مجموعه تعلیمی بدون برچسب را نشان می دهد. شکل 2 نرخ مثبت واقعی (true positive) (TPR) و نرخ مثبت غیر واقعی (FPR) متعلق به شناسایی نمونه روز صفر ایجاد شده توسط کشف ناشناس را گزارش می دهد.



شکل 2 تاثیرات پارامترها بر کشف ناشناس َa) TPR و FPR برای تعداد مختلف از خوشه های k b) TPR و FPR برای اندازه های مختلف تعلیم بدون برچسب

برپایی(setup) آزمایشی که ما مورد استفاده قرار دادیم با آنکه در بخش V-B استفاده کردیم بدون تناقض است. TPR نرخ مجموع ترافیک روز صفر به درستی شناخته شده است نسبت به مجموع تمام ترافیک روز صفر واقعی. FPR نرخ مجموع ترافیک نادرست تشخیص داده شده به عنوان روز صفرم نسبت به مجموع ترافیک کاربردهای شناخته شده است. شکل 2 (a) نتایج را با یک Tu =30000 ثابت و k متغیر نشان می دهد. روشن است که در حالی که FPR ایجاد شده در مرحله اول پایین بود، TPR متناسب با آن هم بالا نبود. دومین مرحله به صورت قابل توجهی TPR را بهتر می کند و در ادامه باعث کاهش FRP می شود. TPR متعلق به کشف ناشناس هنگامی که k از 100 تا 4000 تغییر می کرد از حدود 28% تا 99% تغییر می کند. در همین حال FPR آن از 0% تا 20% افزایش پیدا کرده است. کارایی دسته بندی نهای اختلاف بسیار زیادی پیدا می کند اگر k به مقدار زیادی تغییر کند. ضروری است که یک k را برای متعادل ساختن TPR و FPR برای دستیابی به دقت بالای دسته بندی انتخاب کرد. با ثابت کردن k به 1000 و تغییر دادن Tu از 3000 تا 30000 ما شکل 2(b) را به دست می آوریم. این شکل نشان می دهد که افزایش Tu می تواند اندکی بر TPR و FPR تاثیر بگذارد. در مقایسه با اولین مرحله، مرحله ی دوم می تواند به صورت موثری TPR را در حدود 20% افزایش دهد. اگر ما Tu را در کاربرد عملی خارج از کنترل در نظر بگیریم، تنظیمات پارامتر ما بر k تمرکز می کند.

ما یک روش بهینه سازی جدید را که شامل اعتبار سنجی متقابل 10 مرحله ای (10-fold cross validation) و درخت جستجوی دودویی برای یافتن یک k بهینه است را پیشنهاد می دهیم. مزیت روش بهینه سازی دو بخش است: دقت و سرعت. این روش در شمای RTC پیشنهاد شده برای ارزیابی عملکرد همان گونه که در بخش V\_B اشاره شد مورد استفاده قرار گرفته است. در اعتبار سنجی متقابل 10 مرحله ای، مجموعه ی تعلیمی اصلی شامل جریان های ترافیکی برچسبدار و بدون برچسب به صورت تصادفی به ده زیر مجموعه ی با اندازه برابر تقسیم می شود. از بین ده زیرمجموعه یک زیرمجموعه ی تک به عنوان داده های تعلیمی برای آزمایش مدل کشف ناشناس حفظ می شود. بقیه ی نه زیر مجموعه به عنوان داده های تعلیمی استفاده می شوند. فرایند اعتبارسنجی متقابل سپس ده بار تکرار می شود، دقیقا یکبار با هر کدام از ده زیر مجموعه استفاده شده به عنوان داده ی اعتبار سنجی. ده نتیجه ی گرفته شده از مراحل سپس معدل گیری می شوند تا یک تخمین یکتا را بسازد.

یک مسئله جدید که کدام متریک می تواند برای ارزیابی نتایج کشف ناشناس در اعتبار سنجی متقابل استفاده شود در دقت بهینه سازی حیاتی است. دقت یک ارزش یکتای مشترک است که برای اندازه گیری کارایی کلی دسته بندی ترافیک استفاده می شود. هرچند دقت محاسبه شده با استفاده از داه های تعلیمی برچسبدار برای کلاس های شناخته شده نمی تواند کارایی تشخیص ترافیک روز اول را اندازه گیری کند. بر اساس نتایج تجربی به آن صورت که در شکل 2(a) نشان داده شده، ما می یابیم که FPR یک معیار خوب برای اعتبارسنجی متقابل است. از یک دیدگاه نظری، ایده اصلی به این صورت است:

* برای جستجوی برای یک k حداکثر که مثبتهای نادرست را ایجاد نمی کند.

این به توانایی ما برای تشخیص هر تعداد ممکن از نمونه های روز صفر دقیق بدون ایجاد هیچ خطایی اشاره می کند. هرچند نتایج آزمایشی نشان می دهد که TPR به دست آمده از این ایده پایین است. یک مشاهده از شکل 2(a) این است که TRP به صورت قابل توجهی افزایش می یابد اگر FPR به صورت اندک از مقدار صفر افزایش یابد. به صورت عملی آستانه ی مثبت نادرست برای بهینه سازی پارامترها می تواند یک مقدار کم تنظیم شود. بر اساس آزمایشات ما یافتیم که 3% یک مقدار خوب برای FPR است.

یک مشکل این است که جستجو برای یک k بهینه زمان بر است. برای نمونه اگر مجموعه ی تعلیمی 10000 جریان داشته باشد، k ممکن است از 1 تا 10000تغییر کند. خوشبختانه ما یافتیم که FPR یکنواخت است و با افزایش k، افزایش می یابد. بنابراین یک جستجوی دودویی برای یافتن سریع k یاری رسان است و FPRمنطبق به 3% نزدیکتر است. الگوریتم 4 فرایند انتخاب اتوماتیک پارامتر را توضیح می دهد. یک جستجوی دودویی k، زمان لوگاریتمی می گیرد و بسیار به صرفه است. شکل 3 نتایج این روش هوشمند را برای Tu های متفاوت نشان می دهد. مشخص است که یک k بد می تواند به شدت دقت دسته بندی را تحت تاثیر قرار دهد. این روش بهینه سازی می تواند با موفقیت یک k بهینه را بیابد و دقت دسته بندی ترافیکی بسیار خوبی را ایجاد کند. مجموعه داده ای ترافیک استفاده شده در این بخش همچنین به بخش V اشاره می کند.