ارائه روشی آماری جهت دسته‌بندی جریان‌های ترافیکی اسکایپ

نسرین باتمانی[1]، نورالدین پرندین[2]

تاریخ دریافت: ۱۳۹۶/۱۰/۰۲ تاریخ پذیرش: ۱۳۹۷/۰۳/۲۳

چکیده: اسکایپ یکی از قدرتمندترین و با کیفیت‌ترین ابزارهای چت است که به کاربرانش اجازه می‌دهد از سرویس‌های متعددی مانند: انتقال صدا روی پروتکل اینترنت، ارسال پیام فوری، انتقال و اشتراک‌گذاری فایل، ویدیو کنفرانس و پیام صوتی به صورت رایگان بهره ببرند. اسکایپ به دلیل ارائه سرویس‌هایی با کیفیت بالا و هزینه کم، همچنین به دلیل حفظ امنیت کاربران از شهرت و محبوبیت بالایی برخوردار است. به طوری که ترافیک‌های اسکایپ حجم زیادی از ترافیک‌های اینترنت را به خود اختصاص داده است. از این رو مراکز ارائه‌دهنده خدمات اینترنتی به منظور اعمال کیفیت سرویس و مدیریت شبکه نیاز به شناسایی ترافیک‌ها دارند. از طرف دیگر توسعه‌دهندگان اسکایپ به دلیل حفظ امنیت کاربران اقدام به رمزنگاری ترافیک اسکایپ نموده‌اند به طوری که با استفاده از روش‌های قدیمی مبتنی بر پورت و بازرسی عمیق بسته‌ها نمی‌توان ترافیک اولیه اسکایپ را شناسایی نمود. در نتیجه برای شناسایی این نوع ترافیک‌ها از روش‌های آماری بهره می‌بریم. از این رو در این پژوهش از روش‌های یادگیر ماشین بدون نظارت که یک روش آماری است استفاده می‌کنیم که ترافیک سرویس‌های مختلف اسکایپ را از هم تفکیک کنیم. الگوریتم‌های مورد استفاده در این کار K-Means, EM و Density based است. نتایج نشان می‌دهد که الگوریتم EM کارایی بهتری به نسبت سایر الگوریتم‌ها دارد. همچنین با مقایسه راهکار پیشنهادی خود با کارهای پیشین نتایج حاکی از آن است الگوریتم‌ها جریان دانلود را بهتر از سایر جریان‌ها شناسایی می‌کنند.

واژه‌های کلیدی: ترافیک اینترنت، اسکایپ، شناسایی جریان‌های رمزشده، الگوریتم‌های خوشه‌بندی

---

۱. کارشناس ارشد گروه مهندسی فناوری اطلاعات، دانشکده فنی مهندسی، واحد کرمانشاه، دانشگاه آزاد اسلامی، کرمانشاه، ایران ؛ Email:batmani.nasrin66@gmail.com

۲. استادیار گروه ریاضی، دانشکده علوم پایه، واحد کرمانشاه، دانشگاه آزاد اسلامی، کرمانشاه، ایران ؛ Email : n_parandin@yahoo.com

۱. مقدمه

دسته‌بندی و شناسایی جریان‌های ترافیکی اینترنت در چند سال اخیر اهمیت ویژه‌ای پیدا کرده است به طوری که تعداد زیادی از پژوهشگران به کار در این زمینه علاقه‌مند شده‌اند. حجم وسیعی از مقالات (Finsterbusch, M., et al, 2014) به شناسایی و طبقه‌بندی جریان‌های اینترنت اختصاص دارد که هر کدام یک دسته‌بندی خاص از جریان‌ها ارائه داده‌اند و بیش‌تر از روش‌های سنتی برای دسته‌بندی و شناسایی ترافیک‌ها استفاده نموده‌اند. بنابراین توسعه‌دهندگان نرم‌افزارها هم به منظور حفظ امنیت کاربران ترافیک‌ها را رمز و یا از شماره پورت تصادفی استفاده‌می‌کردند. با توجه به افزایش ترافیک‌های رمزشده تعدادی از پژوهشگران به کار در زمینه شناسایی این ترافیک‌ها پرداختند که دیگر نمی‌توان از روش‌های مبتنی بر بازرسی عمیق بسته‌ها [1] و یا استفاده از شماره پورت این ترافیک‌ها را شناسایی نمود. دراین بین ترافیک‌های رمزشده اسکایپ، به دلیل پشتیبانی از سرویس-های متعدد (چت، دانلود، آپلود، انتقال فایل، Skype out) حایز اهمیت بود زیرا علاوه براینکه همه‌ی این سرویس‌ها برای ارسال ترافیک از یک شماره پورت استفاده می‌کنند، آن شماره پورت نیز به صورت تصادفی به هر کاربر تخصیص داده می‌شود. از آنجا که تنها (Korczyński, M. and A. Duda , 2012) به دسته‌بندی جریان‌های ترافیکی اسکایپ پرداخته‌است، ما هم در این پژوهش با بهره‌گیری از روش‌های آماری اقدام به طبقه‌بندی جریان‌های ترافیکی اسکایپ می‌نماییم تا بتوانیم بهبودی در نتایج مقاله ذکر شده حاصل کنیم و با دقت بالاتری جریان‌های اسکایپ را از هم تفکیک کنیم.

این پژوهش شامل شش بخش است. اولین بخش مبانی نظری تحقیق و ضرورت انجام آن را بیان می‌کند. دومین بخش هم بر کارهای مرتبط مروری خواهیم داشت. بخش سوم هم روش انجام تحقیق را ارائه می‌دهیم و در بخش بعد هم نتایج تجربی کار را ارائه می‌دهیم. بخش‌های آخر هم شامل نتیجه‌گیری کار و معرفی مراجع و منابع استفاده شده‌است.

1 Deep Packet Inspection (DPI)

۲. مبانی نظری

شناسایی صحیح جریان‌های ترافیکی اینترنت نقش مهمی در مدیریت ترافیک شبکه، ارائه کیفیت سرویس و یا تشخیص نفوذ دارد. از این‌رو اسکایپ که با توجه به جدول ۱ که سهم زیادی از تماس‌های تلفن را در کل جهان در اختیار دارد، توجه پژوهشگران و محققین را به خود جلب کرده‌است. به طوری که پژوهشگران به دلایل مختلف تلاش کرده اند جریان‌های اسکایپ را از سایر جریان‌ها تفکیک کنند. آن‌طور که در (Korczyński, M. and A. Duda ,2012) آمده‌است نویسندگان مقاله توانسته‌اند به طور صددرصد ترافیک‌های اسکایپ را از غیراسکایپ شناسایی کنند. بنابراین مطالعات اخیر در زمینه شناسایی جریان‌های ترافیکی اسکایپ، از یکدیگر است.

جدول ۱. سهم اسکایپ در بازار تماس‌های بین‌المللی

سال	درصد نفوذ
۲۰۰۵	۲/۹
۲۰۰۶	۴/۴
۲۰۰۸	۸
۲۰۰۹	۱۲
۲۰۱۰	۱۳
۲۰۱۲	۳۳
۱۰۱۳	۳۶
۲۰۱۴	۴۰

۳.پیشینه‌ی تحقیق

با وجود سیل عظیم ترافیک نرم‌افزارهای جدید رمزشده با چالشی در شناسایی جریان‌ها روبه‌رو هستیم. از جمله این‌که هیچ کدام از روش‌ها نمی‌توانند به صورت برخط جریان‌های رمزشده را شناسایی کنند. راه‌های مختلفی برای شناسایی جریان‌های ترافیکی وجود دارد. اولین راه‌حلی که برای شناسایی جریان‌های ترافیکی ارائه شده مبتنی بر پورت است اما از آنجا که اسکایپ هنگام نصب، یک شماره پورت تصادفی به هر کاربر تخصیص می‌دهد و برای ارتباطات از این شماره پورت استفاده می‌کند، بنابراین از این روش نمی‌توان استفاده کرد.

راه‌حل بعدی استفاده از اطلاعات محتوی بسته است، در این روش نیاز به داشتن الگـوی مشـخص برای هر پروتکل است که همیشه دردسترس نیست. در مورد اسکایپ، چون کل محتوی بسته‌هایش را رمز می‌کند این روش هم کمتر اسـتفاده مـی‌شـود. یـوان و همکـاران (Yuan, Z., et al, 2014) بـا استفاده از روش بازرسی عمیق محتوی بسته‌های کنترلی رمز نشـده‌ی آغـازین، الگـوی خاصـی بـرای شناسایی جریان‌های اسکایپ ارائه نمودند و براساس آن سیستم برخط Skytracer را طراحی کردنـد. دقت و فراخوانی این سیستم به ترتیب ۹۸,۹۳٪ و ۹۹,۵۴٪ است.

راهکار بعدی که اخیرا مورد توجه بسیاری از محققین قرار گرفته‌است استفاده از روش‌های آماری و یادگیر ماشین است. روش‌های یادگیر از ویژگی‌های جریان یا بسته‌های آن مانند: طول جریان، توزیع اندازه‌ی بسته‌ها استفاده می‌کنند، از آنجا که این روش‌ها نیازی به محتوی بسته‌ها ندارد برای جریان‌های ترافیکی رمزشده روش مناسبی است.

در مقالـه (Alshammari, R., & Zincir-Heywood, 2009) بـرای تشـخیص ترافیـک رمزشده اسکایپ و SSH از سایر جریان‌های ترافیکی، از روش‌های یادگیری ماشین بهره برده‌است، به این منظور پـنج الگـوریتم C4.5، AdaBoost، RIPPER، SVM و Naïve Bayesian را روی مجموعه داده‌ها اجرا کرده‌اند. این مطالعه نشان می‌دهد کـه الگـوریتم C4.5 در شـبکه‌هـای مختلـف نسبت به سایر الگوریتم‌های مورد ارزیابی نتایج بهتری را بدست می‌آورد.

Branch و همکـاران در مقالـه (Branch, P. A., Heyde, A., & Armitage, G. J, 2009) با کمک گرفتن از روش‌های یادگیر ماشین و با داشتن تنهـا ۵ ثانیـه از هرجریـان، در تشـخیص ترافیک اسکایپ به دقتی معادل ۹۸٪ و فراخونی ۹۹٪ رسیدند. الگـوریتم مـورد اسـتفاده در ایـن کـار C4.5 است که در آن از ویژگی‌هایی مانند: فاصله زمانی بین دو بسته متوالی، تعداد رخـداد بسـته‌هـای کمتر از ۸۰ بایت و ویژگی‌های آماری بسته‌های بیش از ۸۰ بایت استفاده شده‌است.

Korczynski و Duda (Korczyński, M., & Duda, 2014) در سال ۲۰۱۴ بر اساس مدل مارکوف مرتبه اول، روشی برای شناسایی و دسته‌بندی جریان‌های ترافیکی رمـزشـده مبتنی بر SSL/TLS ارائه نمودند. در روش مذکور، به ازای هر جریان ترافیکی یـک مـدل مارکوف مرتبه اول ایجاد شده و با مدل‌های مارکوف موجود در پایگـاه داده مقایسـه مـی-شود. در نهایت، نوع ترافیـک ورودی بـر اسـاس میـزان مطابقـت آن بـا مـدل‌هـای موجـود مشخص می‌گردد. از جمله جریان‌هـای مـورد آمـوزش در ایـن کـار جریـان‌هـای ترافیکـی

Twitter، Dropbox، Gadu-Gadu، Amazon، Mbank و Skype است که مبتنی بر SSL/TLS هستند.

از آنجا که جریان‌های p2p رفتار مشابهی دارند انواع مختلفی از این جریان‌ها مانند: PPStream,Skype, BitTorrent, Kougou, eDonkey Xunlei, PPlive برای ساخت طبقه‌بند استفاده کرده‌اند. با توجه به این‌که نتایج خود را در دو حالت از مجموعه داده‌ی یکنواخت که به طور مساوی از همه‌ی جریان‌ها استفاده شده و در حالت غیر یکنواخت آزمودند که نتایج تقریبا یکسانی به ازای الگوریتم‌های SVM, Naiive bayes, C4.5, random forest را داشتند و نتایج C4.5 به طور میانگین بهتر از الگوریتم‌های دیگر بوده است. از این‌رو از C4.5 برای ساخت مدل استفاده شد. نتایج این کار نشان می‌دهد که طبقه‌بند بیشترین درصد شناسایی و دقت را در جریان‌های اسکایپ دارد. ۸۴٪ جریان‌های اسکایپ را با دقتی معادل ۹۴٪ شناسایی کرده است.

ابتدا اسکایپ را شناسایی سپس همه‌ی سرویس‌های اسکایپ را از هم تفکیک کرده‌اند. طبق نتایج ارائه شده، F-Measure مربوط به جریان‌های ترافیکی ویدیو و صوت مقادیر ۶۶,۱٪و ۶۴,۲٪ را دارند که به نسبت کمتر از سایر سرویس‌های اسکایپ است. که هدف از این پژوهش بهبود این مقادیر می‌باشد.مقاله‌هایی که در حوزه تشخیص ترافیک اسکایپ ارائه شده‌است، با دقت بالایی جریان‌های اسکایپ را تشخیص داده‌اند. اما تنها مقاله‌ای که در زمینه دسته‌بندی سرویس‌های اسکایپ ارائه شده‌است، مقاله Korczynski و Duda (Korczyński, M. and A. Duda , 2012) است. آن‌ها با استفاده از دو روش بازرسی عمیق بسته‌ها و آماری [1]، طی سه مرحله

٤.مدل تحقیق و روش برآورد

کسب دانش یکی از مهمترین کاربردهای ماشین یادگیرنده است که امروزه بسیار مورد توجه قرار می‌گیرد. به این معنی که عمل یادگیری اطلاعات پایه را از محیط استخراج کرده و برای تحلیل حوادث آینده از آن بهره می‌گیرد. کاربرد یادگیر ماشین، به منظور خوشه‌بندی جریان‌های ترافیکی در روش بدون نظارت به این مقوله بازمی‌گردد. در این

1 SPID

روش جریان‌ها را از نظر شباهت در یک دسته قرار می‌دهیم به طوری که معیار شباهت نزدیکی مقدار ویژگی‌های آن‌ها به همدیگر است. با توجه به نوع الگوریتم خوشه‌بندی، تعداد دسته‌ها را به ورودی الگوریتم می‌دهیم و براساس معیارهایی مانند فاصله‌ی منهتن، فاصله‌ی هر جریان را با خوشه‌ها می‌سنجد و به خوشه‌ای که کمترین فاصله را داشته‌باشد جریان را تخصیص می‌دهد.

خوشه‌بندی در یادگیر ماشین یکی از شاخه‌های یادگیری بی‌نظارت می‌باشد و فرآیندی‌است که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته‌ها خوشه گفته می‌شود. بنابراین خوشه مجموعه‌ای از اشیاء می‌باشد که در آن اشیاء از نظر ویژگی‌هایی با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر غیر مشابه می‌باشند. درمرحله پس از خوشه‌بندی، میزان کارآمدی این الگوریتم‌ها را با معیارهای پرسیژن[1] و ریکال[2] و سپس F-Measure می‌سنجیم. که متداول‌ترین معیارها برای سنجش یک طبقه‌بند می‌باشند (Branch, P. A., Heyde, A., & Armitage, G. J, 2009).

مهمترین بخش در روش‌های یادگیر ماشین، در اختیار داشتن مجموعه داده‌ی مناسب است زیرا ساخت مدل و ارزیابی آن بر اساس این مجموعه داده‌است. هرچه مجموعه داده دارای تنوع بیشتری باشد الگوریتم یادگیر در ارائه مدل پیشنهادی قابلیت اطمینان بیشتری را دارد. در اینجا مراحل ساخت مدل را به ترتیب شرح می‌دهیم.

٤.١.جمع‌آوری مجموعه داده

جریان‌های ترافیکی ما مجموعه‌ای از جریان‌های TCPاسکایپ (۴۲۴ جریان) که مبتنی بر TLS است و جریان‌های ترافیکی بین دو شبکه محلی قرار گرفته در دو کشور فرانسه و لهستان که از طریق شبکه گسترده به هم متصل‌اند با استفاده از نرم‌افزار وایرشارک[3] دریافت شده‌است. همچنین جریان‌های ترافیکی با توجه به شرایط زیر تولید شده‌اند.

- تحت سه سیستم عامل: مک، ویندوز و لینوکس

1 precision
2 recall
3 wireshark

- در دو محیط سیمی و بدون سیم
- جریان‌های ترافیکی بین دو شبکه محلی واقع در فرانسه و لهستان
- نسخه‌های مختلف اسکایپ (۲، ۳، ۵، ۷)

۲.٤. ساخت مجموعه داده ها

از آن‌جا که در ارتباطات TCP یک فاز برقراری ارتباط وجود دارد. به منظور کاهش نویز در مجموعه داده، ده بسته ابتدایی هر جریان را حذف می‌کنیم. سپس مجموعه داده ما که به شکل، فایل‌های خام است. باید به منظور اعمال الگوریتم‌ها روی آن‌ها به فرمت مناسب تبدیل شوند. در این پژوهش، از نرم‌افزار وکا که الگوریتم‌های یادگیر ماشین در آن پیاده‌سازی شده‌است استفاده شده‌است. فرمت متداول نرم‌افزار وکا arrf است،که این فرمت شامل دو بخش اصلی ویژگی[1]و داده[2] است. بخش ویژگی، صفت‌های استخراج شده از مجموعه داده‌ی اولیه است و بخش داده شامل مقادیر ویژگی‌های تعریف شده‌است. تعداد سطرهای فایل برابر با تعداد نمونه‌های ما می‌باشد. به منظور استخراج ویژگی‌های موردنظر از فایل خام از نرم‌افزار نت میت[3] استفاده‌می‌کنیم. نت میت نرم‌افزاری کدباز است که مجموعه داده را به‌عنوان ورودی می‌گیرد و سطر به سطر به ازای هر ویژگی مقادیر آن ویژگی را استخراج می‌کند. از خروجی نت میت با اسکریپت‌نویسی فرمت arrf را که قابل خواندن برای وکا است ایجاد می‌کنیم.

1 attribute
2 data
3 Netmate

```
@ATTRIBUTE std_biat NUMERIC
@ATTRIBUTE duration NUMERIC
@ATTRIBUTE min_active NUMERIC
@ATTRIBUTE mean_active NUMERIC
@ATTRIBUTE max_active NUMERIC
@ATTRIBUTE std_active NUMERIC
@ATTRIBUTE min_idle NUMERIC
@ATTRIBUTE mean_idle NUMERIC
@ATTRIBUTE max_idle NUMERIC
@ATTRIBUTE std_idle NUMERIC
@ATTRIBUTE sflow_fpackets NUMERIC
@ATTRIBUTE sflow_fbytes NUMERIC
@ATTRIBUTE sflow_bpackets NUMERIC
@ATTRIBUTE sflow_bbytes NUMERIC
@ATTRIBUTE fpsh_cnt NUMERIC
@ATTRIBUTE bpsh_cnt NUMERIC
@ATTRIBUTE furg_cnt NUMERIC
@ATTRIBUTE burg_cnt NUMERIC
@ATTRIBUTE total_fhlen NUMERIC
@ATTRIBUTE total_bhlen NUMERIC
@ATTRIBUTE class
{chat,skypeOut,trafficDownload,trafficUpload,video,voice}

@DATA

6,223,26994,266,28763,52,121,1229,137,52,108,1088,121,113,270657
,1008784,174662,103,229040,1046922,162510,113696473,47862,345321
8,8893061,2921943,1061656,3436985,8359722,2226515,13,1587,15,169
1,186,97,0,0,11604,13848,chat
6,302,23050,198,25269,40,76,1071,95,40,127,741,118,9,37182,66568
5,97973,22,56115,510096,88283,12882996,585891,5580538,10575186,7
063498,1721919,1721919,1721919,0,151,11525,99,12634,271,107,0,0,
12088,7928,video
6,274,140064,226,12712,40,511,1500,517,40,56,696,60,42,44228,126
6964,152706,55,53961,609406,115766,13670844,748816,6103227,11457
639,7572281,1464389,1464389,1464389,0,137,70032,113,6356,195,108
,0,0,10968,9276,trafficUpload
```

شکل ۱: ساختار فایل ARRF ایجاد شده

شکل ۱ بخشی از فایل arff را که با اسکریپت‌نویسی ایجاد نمودیم، نشان می‌دهد بخش اول آن مجموعه ویژگی‌ها، آخرین ویژگی برچسب‌های نمونه‌ها است و بخش دوم آن مقادیر آن ویژگی‌هاست. اکنون این فایل قابل پردازش برای وکا است به‌طوریکه می‌توان در نرم‌افزار هر کدام از الگوریتم‌های یادگیر ماشین را انتخاب نمود و طبق آن داده‌ها را خوشه‌بندی کرد.

٤ .٣. اجرای الگوریتم‌های یادگیر ماشین

روش یادگیر ماشین بدون نظارت به این صورت عمل می‌کند که همهٔ الگوریتم‌ها سعی در یافتن بیشترین شباهت را بین نمونه‌های یک خوشه و بیشترین تفاوت را بین نمونه‌های خوشه‌های متفاوت دارند. در این روش از برچسب مجموعه داده فقط به منظور ارزیابی نتایج استفاده‌می‌شود. الگوریتم‌های مورد بررسی K-Means, EM وDensity based است. نحوهٔ کار هر کدام از الگوریتم‌ها را به اختصار شرح می‌دهیم.

الگوریتم K-Means

یکی از روش‌های معتبر خوشه‌بندی، خوشه‌بندی K-means است که یک روش پایه برای سایر الگوریتم‌های خوشه‌بندی محسوب می‌شود. در این الگوریتم بر اساس کمترین فاصله‌های هر داده از مرکز یک خوشه (میانگین) خوشه‌بندی را انجام می‌دهد. برای این الگوریتم شکل‌های مختلفی بیان شده‌است. ولی همهٔ آن‌ها دارای روالی تکراری هستند که برای تعدادی ثابت از خوشه‌ها، سعی در تخمین مقادیری را دارند.

خوشه بندی Density-based

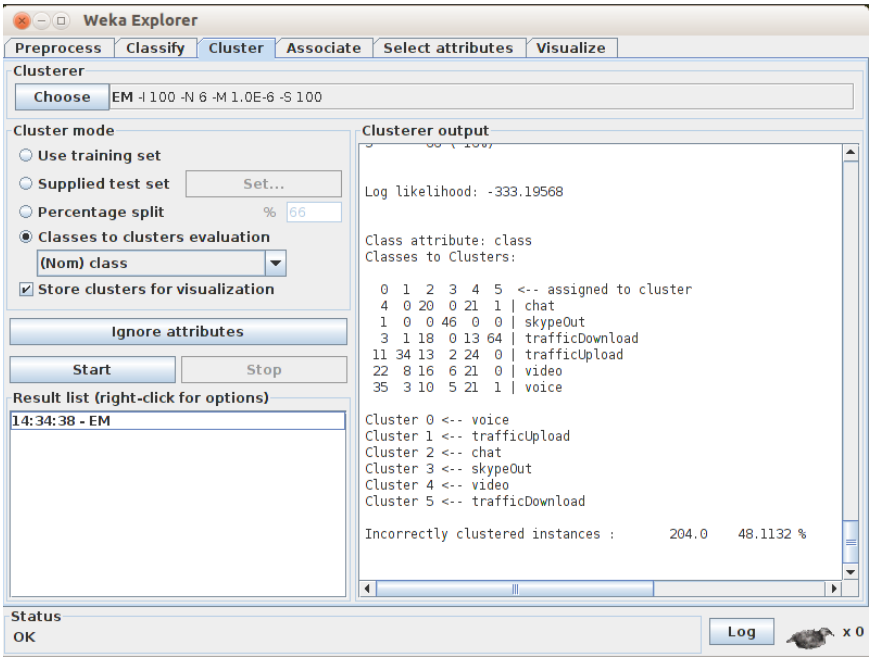
در این تکنیک این اصل مطرح می‌شود که خوشه‌ها مناطقی با چگالی بیشتر هستند که توسط مناطق با چگالی کمتر از هم جدا شده اند. یکی از مهم‌ترین الگوریتم‌ها در این زمینه الگوریتمDBSCAN است. روش این الگوریتم به این صورت است که هر داده متعلق به یک خوشه در دسترس چگالی سایر داده‌های همان خوشه است، ولی در دسترسی چگالی سایر داده‌های خوشه‌های دیگر نیست (چگالی داده همسایگی به مرکز داده و شعاع همسایگی دلخواه ε است). مزیت این روش این است که تعداد خوشه‌ها به صورت خودکار مشخص می‌شود. در تشخیص نویز نیز بسیار کاراست.

الگوریتم EM

الگوریتم امید ریاضی بیشینه‌سازی یک روش تکرارشونده است که به دنبال یافتن برآوردی با بیشترین درست‌نمایی برای پارامترهای یک توزیع پارامتری است. این الگوریتم روش متداول برای زمان‌هایی‌است که برخی از متغیرهای تصادفی پنهان هستند.

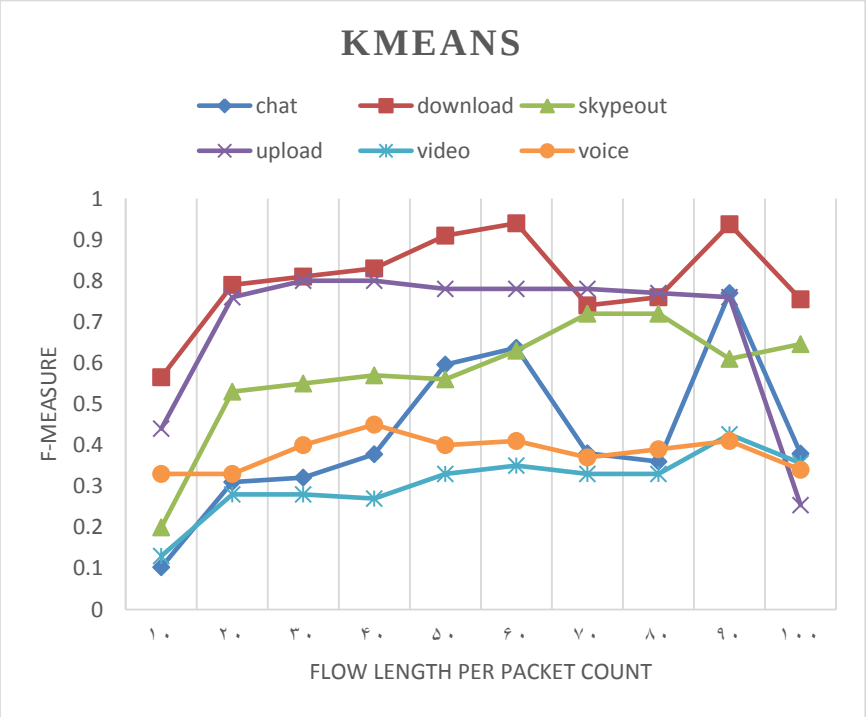
۵.داده‌ها و نتایج تجربی

دانشی که در مرحله یادگیری مدل ارائه می‌شود، باید در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان میزان کارایی آن را تعیین نمود. همچنین لازمه موفقیت در بهره‌مندی از علم داده‌کاوی تفسیر درست دانش ایجاد شده و ارزیابی آن است. به منظور مقایسه روش پیشنهادی با روش‌های پیشین از معیارهای ریکال، پرسیژن و F-Measure استفاده می‌کنیم. در ابتدا به منظور اجرای سریعتر الگوریتم‌ها، مجموعه ویژگی‌های غیرمفید (شماره پورت، درس مبدا، آدرس مقصد، ...) را حذف می‌کنیم. از آن‌جا که شناسایی جریان‌ها نیاز است به صورت بلادرنگ انجام شود از این‌رو، طول جریان را محدود به تعدادی از بسته‌های آغازین نمودیم. به طوری که ابتدا ده بسته ابتدایی که همان مراحل دست تکانی سه مرحله ای است، حذف نمودیم سپس طول جریان را از ده بسته آغاز کردیم و هر بار ده بسته به آن اضافه نمودیم. به ازای هرکدام از جریان‌ها با طول‌های متفاوت الگوریتم‌های انتخابی را اجرا می‌کنیم. پس از انتخاب الگوریتم و تنظیمات مربوط به آن از جمله تعداد خوشه، با انتخاب یکی از حالات چهارگانه، اقدام به اجرای الگوریتم می‌نماییم. در اینجا حالت classes to clusters evaluation اجرا می‌کنیم تا بتوان با استفاده از ماتریس درهم ریختگی که در خروجی تولید می‌گردد میزان کارایی و صحت دسته‌بندی سرویس‌های اسکایپ را ارزیابی کنیم.



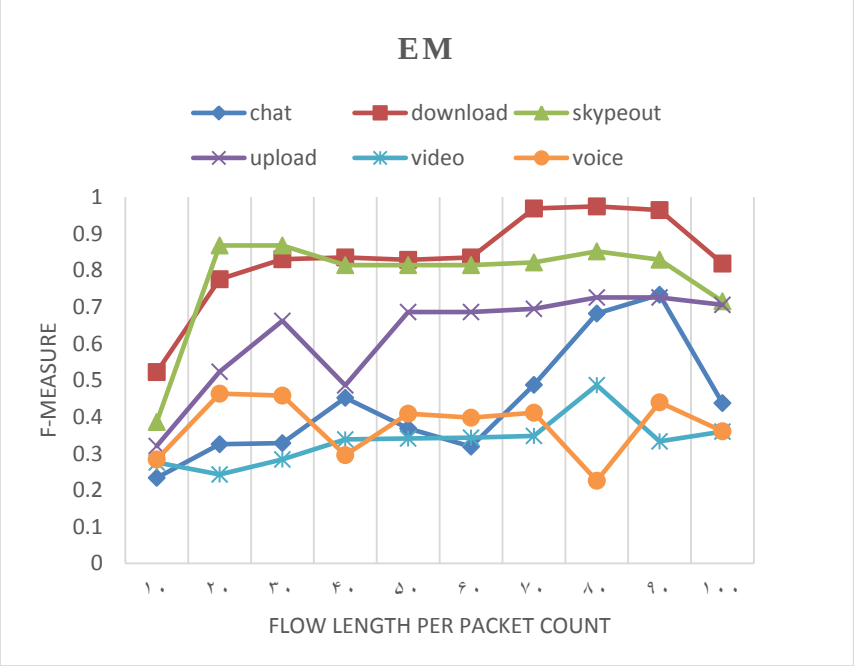
شکل ۲: اجرای الگوریتم EM روی مجموعه داده‌های اسکایپ در نرم‌افزار وکا

شکل ۲ نمایی از اجرای الگوریتم EM را در خوشه‌بندی جریان‌های اسکایپ نشان می‌دهد. از برچسب داده‌ها که به‌عنوان آخرین ویژگی تعریف کردیم فقط در مرحله ارزیابی الگوریتم استفاده می‌شود. خروجی این الگوریتم‌ها شامل ماتریس درهم‌ریختگی است که می‌توان با استفاده از این ماتریس، الگوریتم را از نظر کارایی و دقت مورد بررسی قرار دهیم.



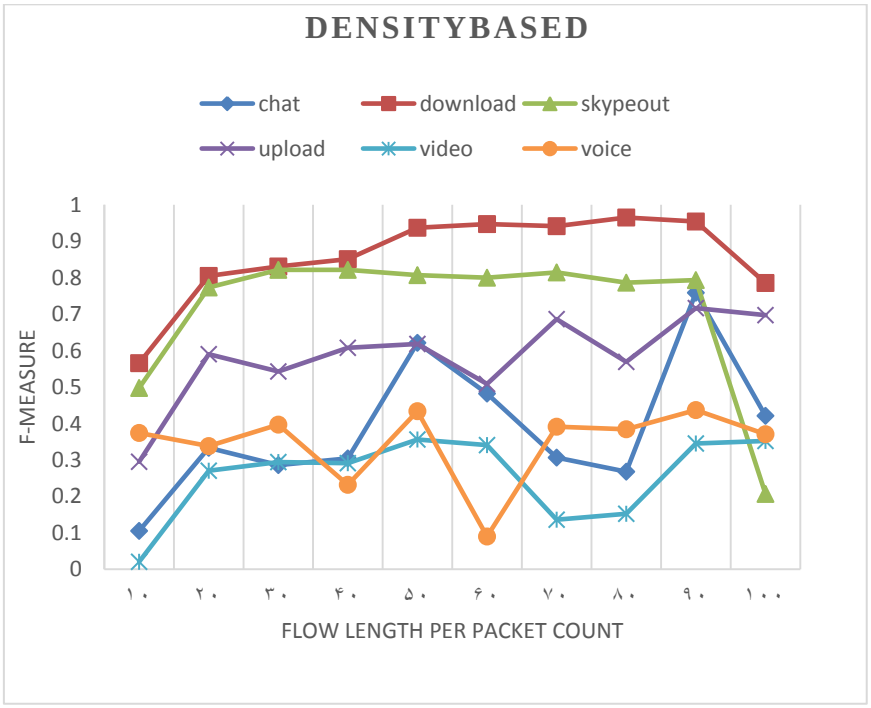
نمودار ۱: مقدار F-Measure مربوط به الگوریتم Kmeans در شناسایی جریان‌های مختلف اسکایپ در جریان‌هایی به طول کمتر از ۱۰۰ بسته حاوی داده

از آنجا که الگوریتم Kmeans یک الگوریتم پایه برای بسیاری از الگوریتم‌ها است، بنابراین در این تحقیق از این الگوریتم، به‌عنوان یکی از الگوریتم‌های خوشه‌بندی استفاده شده‌است. نمودار ۱ مربوط به اجرای الگوریتم Kmeans است، با افزایش تعداد بسته‌های جریان تا ۱۰۰ بسته، الگوریتم با F-Measure بیش از ۹۵ درصد جریان‌های دانلود را شناسایی می‌کند. مشاهده می‌شود جریان‌های چت، آپلود و Skype out هم با افزایش طول جریان مقدار F-Measure آن‌ها نیز افزایش می‌یابد. جریان‌های video و voice به دلیل الگوی رفتاری شبیه به هم مانند: اندازه بسته‌های کوچک و فاصله زمانی کم بین بسته-ها، به خوبی از هم تشخیص داده نمی‌شوند. از این‌رو افزایش طول جریان، تاثیر به‌سزایی در شناسایی این نوع جریان‌ها ندارد.



نمودار ۲: مقدار F-Measure مربوط به الگوریتم EM در شناسایی جریان‌های مختلف اسکایپ در جریان‌هایی به طول کمتر از ۱۰۰ بسته حاوی داده

طبق نمودار ۲ الگوریتم EM جریان‌های download و Skype out را بهتر از سایر جریان‌ها شناسایی نموده است. جریان‌های download به دلیل اندازه بسته‌های بزرگ ارسالی و جریان‌های skype out هم به دلیل اندازه بسته‌های با طول تقریبا یکسان تفکیک کننده‌ی خوبی از سایر جریان‌ها می‌باشند. جریان‌های آپلود و چت نیز به نسبت video و voice مقدار F-Measure بالاتری را دارند. از آنجا که الگوریتم EM نسبت به داده‌های گمشده حساس نیست فراز و نشیب‌های کمتری در این نمودار نسبت به Kmeans مشاهده می‌شود.



نمودار ۳: مقدار F-Measure مربوط به الگوریتمDensity based در شناسایی جریان‌های مختلف اسکایپ در جریان‌هایی به طول کمتر از ۱۰۰ بسته حاوی داده

این الگوریتم هم تقریبا مانند EM عمل کرده‌است اما فراز و نشیب‌های بیشتری نسبت به EM دارد. این فراز و نشیب حاکی از آن است با افزایش طول جریان به ازای chat, voice و video الگوی مشخصی با توجه به ویژگی‌های تعریف شده‌ی آن‌ها استخراج نمی‌شود. الگوریتم به دلیل داشتن ویژگی‌های مشخص (اندازه بسته‌های تقریبا ثابت) آپلود, دانلود و Skype out را بهتر از سایر جریان‌ها شناسایی کرده‌است. جریان‌های chat به دلیل نرسیدن به الگویی ثابت در جریان‌هایش (اندازه بسته‌های متفاوت، فاصله‌های متفاوت بین دو بسته متوالی و ...)، افزایش طول جریان به طور قطع سبب شناسایی بهتر آن نمی‌گردد.

جدول۲. مقایسه راهکار پیشنهادی با روش پیشنهاد شده توسط دیگر محققین (Korczyński, M. and A. Duda , 2012)

	نتایج حاصل از روش آماری در جریان‌هایی به طول ۶۰۰ بسته آغازین (Korczyński, M. and A. Duda , 2012)			نتایج حاصل از خوشه بندی EM در جریان‌هایی به طول ۹۰ بسته آغازین (راهکار پیشنهادی)		
نوع سرویس	پرسیژن%	ریکال%	F-M%	پرسیژن%	ریکال%	F-M%
Chat	۹۰/۹	۹۳	۹۲	۶۰	۹۶	۷۳
Skypeout	۱۰۰	۹۰/۳	۹۴/۹	۷۱	۹۸	۸۲
Download	۱۰۰	۹۲/۹	۹۶/۳	۹۷	۹۶	۹۳
Upload	۹۶/۹	۹۰	۹۳/۳	۹۶	۵۸	۷۲
Video	۶۳	۷۵/۴	۶۸/۷	۳۴	۳۳	۳۳
Voice	۷۴	۵۶/۱	۶۳/۸	۴۷	۴۱	۴۴

نتایج حاصل از راهکار پیشنهادی مبتنی بر خوشه‌بندی نشان می‌دهد از بین سه الگوریتم اجرا شده، الگوریتم EM نسبت به دو الگوریتم دیگر نتایج بهتری را داشته‌است. حال نتایج کار خود را با راهکار آماری مقایسه می‌کنیم. در راهکار پیشنهادی از جریان‌هایی به طول ۹۰ بسته و در روش آماری از جریان‌هایی به طول ۶۰۰ بسته استفاده می‌کنیم. با توجه به جدول ۲ مشاهده می‌شود راهکار آماری مقادیر F-Measure بالاتری نسبت به راهکار پیشنهادی دارد. اما نکته حایز اهمیت این است که بتوان با تعداد بسته‌های کمتر یعنی در کمترین زمان ممکن جریان راشناسایی نمود.

٦.نتیجه‌گیری

از آنجا که دسته‌بندی و شناسایی جریان‌های اینترنتی اهمیت فراوانی پیدا کرده‌است، در این مقاله روشی، به منظور دسته‌بندی جریان‌های رمزشده اسکایپ ارائه شده‌است. نتایج حاکی از آن است با توجه به نوع الگوریتم‌های انتخابی و مجموعه ویژگی‌های انتخاب شده، روش‌های خوشه‌بندی، روش مناسبی برای شناسایی و دسته‌بندی ترافیک‌ها نمی‌باشند، شاید بتوان با تحقیق و بررسی و در نظر گرفتن مجموعه ویژگی مناسب و الگوریتم‌هایی دیگر این نتایج را بهبود داد. از جمله ویژگی که در این کار به آن توجهی نشده‌است می‌توان به

این اشاره نمود که در ارسال صوت هر لحظه فقط یک نفر در حال ارسال ترافیک می‌باشد اما در جریان‌های ترافیکی ویدیو هر لحظه تقریبا دو طرف به یک اندازه ترافیک ارسال می‌کنند که می‌تواند وجه تمایزی بین جریان‌های ترافیکی ویدیو و صوت باشد. امید است در کارهای آتی بتوان با بهره بردن از الگوریتم‌های دیگر یادگیر ماشین میزان F-Measure را به ازای همه‌ی جریان‌ها به ویژه video و voice افزایش داد.

منابع

1- Finsterbusch, M., Richter, C., Rocha, E., Muller, J. A., & Hanssgen, K. (2014). A survey of payload-based traffic classification approaches. IEEE Communications Surveys & Tutorials, 16(2), 1135-1156.

2- Korczyński, M., & Duda, A. (2012, June). Classifying service flows in the encrypted skype traffic. In Communications (ICC), 2012 IEEE International Conference on (pp. 1064-1068). IEEE.

3- Yuan, Z., Du, C., Chen, X., Wang, D., & Xue, Y. (2014, February). Skytracer: Towards fine-grained identification for skype traffic via sequence signatures. In Computing, Networking and Communications (ICNC), 2014 International Conference on (pp. 1-5). IEEE.

4- Alshammari, R., & Zincir-Heywood, A. N. (2009, July). Machine learning based encrypted traffic classification: Identifying ssh and skype. In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on (pp. 1-8). IEEE.

5- Branch, P. A., Heyde, A., & Armitage, G. J. (2009, June). Rapid identification of Skype traffic flows. In Proceedings of the 18th international workshop on Network and operating systems support for digital audio and video (pp. 91-96). ACM

6- Korczyński, M., & Duda, A. (2014, April). Markov chain fingerprinting to classify encrypted traffic. In Infocom, 2014 Proceedings IEEE (pp. 781-789). IEEE.