دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش درس روش پژوهش و ارائه

طبقه بندی ترافیک شبکه
با استفاده از الگوریتم‌های یادگیری ماشین

نگارش

محمدمهدی هجرتی

استاد راهنما

دکتر رضا صفابخش

اردیبهشت ۱۴۰۰

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش درس روش پژوهش و ارائه

طبقه بندی ترافیک شبکه
با استفاده از الگوریتم‌های یادگیری ماشین

نگارش

محمدمهدی هجرتی

استاد راهنما

دکتر رضا صفابخش

اردیبهشت ۱۴۰۰

سپاس گزاری

بی‌تردید تهیه‌ی این گزارش بدون راهنمایی‌های ارزشمند استاد بزرگوار جناب آقای دکتر رضا صفابخش میسر نمی شد. بدین وسیله بر خود لازم می‌دانم از زحمات بی‌دریغ ایشان صمیمانه تقدیر و تشکر نمایم.

محمدمهدی هجرتی

اردیبهشت ۱۴۰۰

چکیده

امروزه با توجه به استفاده‌ی روزافزون از شبکه‌ی اینترنت در دنیا، افزایش سریع تعداد کاربران و ظهور برنامه‌های کاربردی تحت شبکه، ترافیک اینترنت به شدت در حال افزایش است. در نتیجه شناسایی برنامه‌ها در شبکه، به امر پیچیده‌ای تبدیل شده است. از طرفی طبقه‌بندی جریان‌ها نقش مهمی در امنیت و مدیریت شبکه و به‌ویژه برای مقابله با حملات دارد. در گذشته از روش‌های گوناگونی برای طبقه‌بندی ترافیک اینترنت از جمله روش‌های مبتنی بر درگاه، یا بررسی پیلود بسته‌ها استفاده می‌شد. اما امروزه با توجه به مشکلات و محدودیت‌های موجود در روش‌های قبل مثل اختصاص دادن درگاه به صورت پویا، وجود داده های رمزگذاری شده و ... ناگزیر مجبور به استفاده از روش‌های جدید مثل یادگیری ماشین شده‌ایم. روش‌ها و الگوریتم‌های متعددی با استفاده از یادگیری ماشین برای طبقه‌بندی ترافیک شبکه پیشنهاد شده‌است. هدف این پژوهش بررسی این روش‌ها و ارزیابی و مقایسه‌ی روش‌های پیشنهادی موجود می‌باشد.

واژه‌های کلیدی:

ترافیک شبکه، یادگیری ماشین، طبقه‌بندی، دسته‌بندی

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

فصل اول

مقدمه

مقدمه

امروزه طبقه بندی ترافیک شبکه به یک موضوع مهم در حوزه‌ی کامپیوتر تبدیل شده‌است. برای ارائه‌دهندگان خدمات اینترنت[1]، آگاهی از برنامه‌های اجرا شده در شبکه یک امر حیاتی می‌باشد. طبقه‌بندی ترافیک شبکه اولین مرحله برای تجزیه و تحلیل و شناسایی انواع مختلف برنامه‌های شبکه است. با این روش ارائه‌دهندگان خدمات اینترنتی یا اپراتورهای شبکه می‌توانند عملکرد کلی یک شبکه را مدیریت کنند.

روش های اولیه ی طبقه بندی ترافیک اینترنت مبتنی بر بازررسی بسته[2] های جریان بودند. روش مبتنی بر شماره درگاه[3]، شماره درگاه در سرآیند[4]، بسته ها را با شماره های درگاه ثبت شده در مرجع شماره های اختصاص داده شده اینترنت مقایسه می کند. این روش برای جریان های با شماره درگاه پویا قابل اجرا نیست.[۱]

روش طبقه بندی مبتنی بر پیلود[5]، تشخیص نوع برنامه را با پیدا کردن برخی از ویژگی های منحصر به فرد برنامه ها انجام می دهد. این روش روی بازرسی داده های کاربر تکیه دارد و درنتیجه، باعث نقص حریم خصوصی کاربر می شود.

روش مبتنی بر رفتار میزبان، مستقل از بازرسی بسته های جریان، با نظارت بر همه جریان های ارسالی یا دریافتی روی میزبان های شبکه، می تواند ترافیک ایجاد شده توسط برنامه ها را طبقه بندی کند. این روش مبتنی بر این فرض است که میزبان در هر لحظه یک برنامه را اجرا می کند. که در واقعیت معمولا این طور نیست.

امروزه متداول‌ترین تکنیک مورد استفاده، یادگیری ماشین[6] است، که توسط بسیاری از محققان استفاده می‌شود و باعث بدست آمدن نتایج به مراتب دقیق‌تری از روش‌های پیشین شده است. تکنیک‌های یادگیری ماشین، با استفاده از مجموعه ویژگی های آماری جریان به طور خودکار الگوهای ساختاری موجود در انتقال داده‌های جریان را کشف می کنند. این روش می تواند مشکلاتی مانند شماره درگاه پویا، عدم حفظ حریم خصوصی کاربران و فرض عدم اجرای همزمان چند برنامه روی یک میزبان را رفع نماید. هدف از این پژوهش بررسی این روش ها و ارزیابی و مقایسه‌ی روش‌های پیشنهادی موجود

[1] Internet service providers
[2] packet
[3] port
[4] header
[5] payload
[6] machine learning

می‌باشد.

در ادامه، در فصل دوم روش‌های مختلف موجود برای طبقه‌بندی ترافیک شبکه و مشکلات موجود بیان شده‌است. در فصل سوم مدل‌سازی و روش پیاده‌سازی روش‌های مبتنی بر یادگیری ماشین مورد بررسی قرار گرفته‌است. و در نهایت جمع‌بندی، نتیجه‌گیری و پیشنهادها در فصل پنجم ارائه شده‌است.

فصل دوم

انواع روش‌های طبقه‌بندی ترافیک شبکه

از گذشته روش‌های مختلفی برای طبقه‌بندی ترافیک شبکه اینترنت وجود داشته‌است که در این بخش هر یک از این تکنیک‌ها را بررسی می کنیم.

۲-۱ روش مبتنی بر درگاه

شناخته شده ترین و قدیمی ترین روش مورد استفاده برای طبقه بندی ترافیک اینترنت، تطبیق شماره ی درگاه است. در این روش، از شماره‌ی درگاه مقصد در سرآیند لایه انتقال بسته برای شناسایی ترافیک استفاده می‌شود و مقدار شماره‌ی درگاه با لیست شماره درگاه های تعیین شده در استاندارد IANA، برای شناسایی بسته جاری مقایسه می‌شود. در جدول ۱ شماره درگاه اختصاص داده شده برای بعضی از برنامه‌های معروف آورده شده‌است. مثلا برنامه‌های وب از پورت ۸۰ استفاده می‌کنند. [۱]

جدول ۲-۱: شماره درگاه‌های اختصاص داده شده به برخی از برنامه‌های پرکاربرد

برنامه	شماره درگاه اختصاص داده شده
FTP Data	۲۰
FTP	۲۱
SSH	۲۲
Telnet	۲۳
SMTP	۲۵
DNS	۵۳
HTTP	۸۰
POP3	۱۱۰
NTP	۱۲۳

اما امروزه برنامه‌های جدید بخصوص برنامه‌های نظیر به نظیر از روش‌های مختلف برای پنهان کردن خود استفاده می‌کنند. آنها از درگاه‌های پویا و یا درگاه‌های دیگر برنامه‌های شناخته‌شده در اتصالاتشان استفاده می‌کنند. که این امر باعث کاهش دقت طبقه‌بندی این روش شده‌است.

۲-۲ روش مبتنی پیلود

با بوجود آمدن برنامه های نظیر به نظیر این روش، جایگزین قبلی شد. در این روش محتوای بسته‌ها برای پیدا کردن امضای برنامه‌های شناخته شده جستجو می‌شود. در جدول ۲ یک نمونه از این امضاها که توسط کاراگیانیس[1] استفاده شده است را می‌بینیم.

جدول ۲-۲: نمونه امضاهای موجود در بسته های برخی از برنامه های پرکاربرد[۳]

P2P Protocol	String	Trans. Protocol
Edonkey 2000	0xe319010000	TCP/UDP
	0xe53f010000	
Fasttrack	"Get /.hash"	TCP
	0x2700000002980	UDP
BitTorrent	"0x13Bit"	TCP
Gnutella	"GNUT" "GIV"	TCP
Aress	"GET hash"	UDP
	"Get Shal"	

این روش به نسبت روش قبل دقیق تر عمل می‌کند اما چند مشکل دارد. مهم تر از همه این که نمی توان آن را بر روی بسته های رمزگذاری شده اعمال کرد. به علاوه تجزیه و تحلیل مستقیم داده ها باعث نقض حریم خصوصی کاربران می‌شود. همچنین این روش چون محتویات تمام بسته ها را بررسی میکند، نیازمند سیستم پردازشی به مراتب قوی تری نسبت به روش های دیگر است.

۲-۳ روش مبتنی بر رفتار میزبان

ایده ی اصلی این روش این است که برنامه های مختلف الگوهای اتصال متفاوتی دارند. روش مبتنی بر رفتار میزبان، ارتباطات میان میزبان های خاص در یک شبکه را شناسایی میکند و الگوی ارتباط یک میزبان خاص با الگوی رفتار فعالیت های متفاوت مقایسه می‌شود. این روش پیلود بسته را برای طبقه بندی ترافیک استفاده نمی‌کند. پس می‌تواند بسته ها با محتوای رمزگذاری شده را نیز شناسایی کند. اما اشکال این روش این است که بیشتر تکنیک های مبتنی بر رفتار میزبان فرض می‌کند که میزبان

[1] Thomas Karagiannis

های تحت نظارت در یک لحظه تنها از یک برنامه استفاده می‌کنند اما می‌دانیم در واقعیت، این وضعیت ممکن است هرگز اتفاق نیفتد. بیشتر کاربران از برنامه های زیادی به طور همزمان استفاده می‌کنند. مشکل دیکری که در این روش بوجود میاد این است که برخی الگوهای رفتاری برنامه را نمی‌توان به آسانی کشف کرد. به مقدار حافظه و جریان های زیادی برای همه میزبان ها نیاز دارد تا بتواند الگوی اتصال را کشف کند. [۴]

۲-۴ روش مبتنی بر یادگیری ماشین

یادگیری ماشین مجموعه ای از تکنیک‌ها برای داده کاوی و کشف دانش است که الگوهای ساختاری مفید در داده ها را جستجو میکند. این روش طبقه بندی مبتنی بر مجموعه داده های برچسب گذاری شده است. در این روش، یک طبقه بندی کننده یادگیری ماشین به عنوان ورودی آموزش داده می‌شود و سپس با استفاده از نمونه آموزش دیده، داده های ناشناخته طبقه بندی می‌شود. دو روش اصلی در یادگیری ماشین وجود دارد. که در ادامه به بررسی هرکدام می‌پردازیم.

۲-۴-۱ یادگیری نظارت شده

روش های یادگیری با نظارت، مبتنی بر دانش از پیش تعریف شده هستند. این الگوریتم ها در مرحله آموزش، نمونه های از پیش طبقه بندی شده (متشکل از ویژگی ها و برچسب مرتبط با آنها) را به عنوان ورودی می‌گیرند و قوانین طبقه بندی ایجاد می‌شود. در مرحله طبقه بندی تلاش می‌کنند تا برچسب نمونه‌های بدون برچسب را پیشبینی کنند.

۲-۴-۲ یادگیری بدون نظارت

روش های یادگیری بدون نظارت، در مرحله یادگیری خود به هیچ دانش از قبل تعیین شده ای نیاز ندارند. این روش گروه های طبیعی را در داده ها کشف می‌کنند و روی کشف الگوهای موجود در داده ها تمرکز دارند. نمونه ها را بر اساس میزان شباهت ویژگی‌های آنها که توسط یک رویکرد اندازه گیری فاصله تعریف می‌شود. بنابر این در خروجی نمی‌توانیم نوع داده را به طور دقیق مشخص کنیم. صرفا داده های مشابه با هم در یک دسته قرار می‌گیرند.

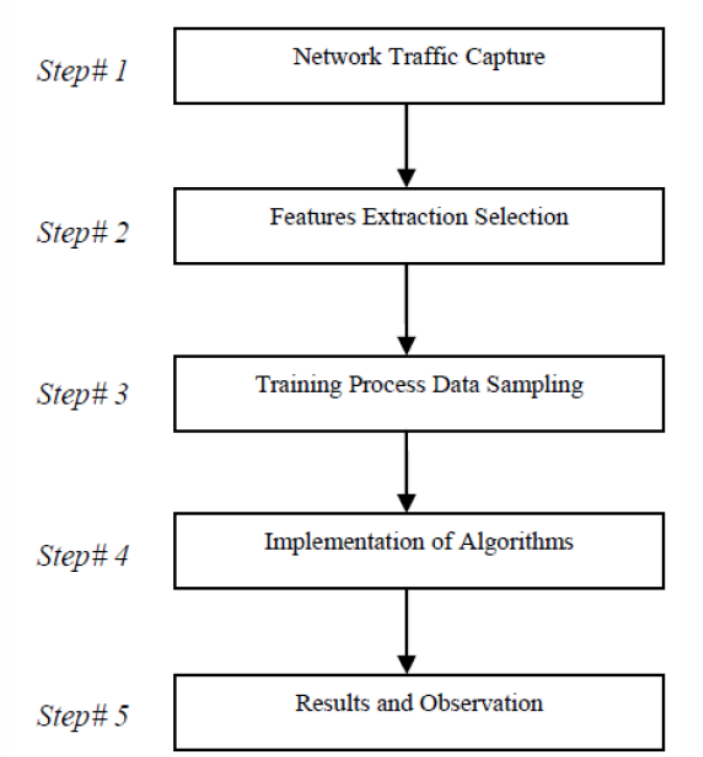
۲-۵ خلاصه

در این فصل روش های کلی موجود برای طبقه بندی ترافیک شبکه از ابتدا تا کنون مورد بررسی قرار گرفت و راجع به مزایا و معیاب هر کدام بحث شد. روش مبتنی بر درگاه در برنامه های نظیر به نظیر دقت بالایی ندارد. روش مبتنی بر پیلود نیازمند سیستم پردازشی قوی می‌باشد و در شناسایی بسته های رمزگاری شده عاجز است. روش مبتنی بر رفتار میزبان صرفا برای زمان هایی که میزبان تنها از یک برنامه در لحظه استفاده می‌کند نتایج دقیقی تولید می‌کند. سپس روش های مبتنی بر یادگیری ماشین و انواع آن معرفی شد. در ادامه روش مدل سازی و استفاده از یادگیری ماشین در طبقه بندی ترافیک شبکه توضیح داده می‌شود.. و سپس چند مورد از الگوریتم های معروف یادگیری ماشین و نیز پژوهش های انجام شده حول این موضوع بررسی می‌شود.

فصل سوم

مدل سازی روش یادگیری ماشین

در این بخش مدل سازی و پیاده سازی انجام طبقه بندی ترافیک با استفاده از یادگیری ماشین در پنج مرحله بررسی می‌شود. شکل زیر خلاصه ای از این مراحل را نشان می‌دهد.



شکل ۳-۱: مدل پنج مرحله ای طبقه بندی ترافیک[۳]

همانطور که مشاهده می‌شود این پنج مرحله شامل جمع آوری داده، استخراج ویژگی ها، یادگیری نمونه، پیاده سازی الگوریتم و تحلیل نتایج می‌باشد، که در ادامه هر مرحله به اختصار توضیح داده می‌شود.

۳-۱ جمع آوری داده

اولین و مهم ترین بخش، جمع آوری داده یا به اصلاح گرفتن بسته[۱] های شبکه می باشد. ابزار های مختلفی برای این کار وجود دارد. از جمله معروف ترین ابزار ها در این زمینه وایرشارک[۲] و تی سی پی دامپ[۳] هستند که در پژوهش های مختلف از هر کدام از این ابزار ها استفاده شده است.

[1] packet capture
[2] Wireshark
[3] tcpdump

۳-۲ استخراج ویژگی ها

این مرحله که در پژوهش های مختلف غالبا با ابزارهای نت میت[۴] و پرل اسکریپت[۵] انجام می‌گیرد، ویژگی های خاصی از بسته ها استخراج می‌شود و با کمک آن ها دسته بند یادگیری ماشین[۶] را آموزش می‌دهند. از جمله ویژگی های قابل استخراج می‌توان تعداد بسته ها، طول هر بسته، درگاه و پروتکل[۷] مورد استفاده و غیره را نام برد.

۳-۳ یادگیری نمونه

در مرحله ی سوم لازم است تا از داده های مورد نظر که در بخش اول بدست آمده، نمونه گیری انجام شود. از طرفی چون در این پژوهش از الگوریتم های یادگیری نظارت شده استفاده شده است، بر روی داده های دریافتی برچسب گذاری نیز انجام می‌شود تا بتوان به کمک آنها، بسته های ناشناخته را طبقه بندی کرد.

۳-۴ پیاده سازی الگوریتم

در این مرحله باید الگوریتم یادگیری ماشین مورد نظر بر روی داده های آموزش داده شده اعمال و پیاده سازی شوند. پژوهشگران معمولا به کمک ابزار وکا[۸] این کار را انجام می‌دهند.

۳-۵ بررسی و تحلیل نتایج

در نهایت در بخش آخر ابراز وکا فرآیند تست روی داده ها را انجام می‌دهد و دقت ارزیابی انجام شده برای الگوریتم های مختلف را برای ما مشخص می‌کند. با بررسی پژوهش های انجام شده بر روی شش الگوریتم یادگیری ماشین درخت تصمیم آر بی اف[۹]، ماشین بردار پشتیبان[۱۰]، C4.5، نزدیک ترین

[4] Netmate
[5] Perl script
[6] machine learning classifier
[7] protocol
[8] Weka classification simulation tools
[9] RBF decision tree
[10] support vector machine (SVM)

همسایه[11]، نیوبیز[12]، و شبکه بیز[13] نتایج زیر حاصل شده است. [۳، ۲]

جدول ۳-۱: دقت طبقه بندی الگوریتم های مختلف یادگیری ماشین

الگوریتم	Naive Bayes	SVM	Bayes Net	RBF	C4.5	NN
دقت طبقه بندی	٪۷۱.۸۹	٪۷۴.۰۵	٪۷۸.۳۲	٪۶۸.۲۵	٪۹۳.۳۳	٪۸۰.۲

همانطور که در جدول نشان داده شده است، از بین الگوریتم های مورد استفاده در این پژوهش، با استفاده از الگوریتم یادگیری ماشین C4.5 توانسته ایم به دقت بیش از ۹۳ درصد برای طبقه بندی ترافیک شبکه دست پیدا کنیم. الگوریتم نزدیک ترین همسایه نیز دقت تقریبا ۸۰ درصدی را نشان می‌دهد. پس از آن برای الگوریتم های شبکه بیز، ماشین بردار پشتیبان و نیو بیز نیز دقتی بین ۷۰ تا ۸۰ درصد بدست آمده است.

۳-۶ خلاصه

در این بخش قدم به قدم با مراحل مدل سازی و پیاده سازی روش یادگیری ماشین آشنا شدیم و مشاهده کردیم که الگوریتم یادگیری ماشین C4.5 با دقتی معادل ۹۳.۵ درصد، بسته های شبکه را به درستی طبقه بندی می‌کند.

[11] nearest neighbor

[12] naive bayes

[13] bayesian network

فصل چهارم

نتیجه گیری و پیشنهادها

۴-۱ نتیجه گیری

شناسایی جریان جاری روی ترافیک اینترنت روی جنبه های مختلف شبکه مانند امنیت تأثیر زیادی دارد. همچنین تشخیص جریان های ترافیک باعث برنامه ریزی صحیح در قسمت های مختلف شبکه مانند تخصیص منابع، بهبود کیفیت خدمات سرویس و غیره می شود. بنابراین، با توجه به اهمیت شناسایی جریان های ترافیک اینترنت، در این پژوهش انواع روش های موجود برای طبقه بندی ترافیک که از گذشته تا کنون استفاده می‌شده است، مورد بررسی قرار گرفت و در مورد مزایا و معایب هرکدام بحث شد. در ادامه آزمایش های انجام شده برای شش مورد از الگوریتم های یادگیری ماشین برای طبقه بندی ترافیک شبکه با یکدیگر مقایسه شد که دیدیم الگوریتم C4.5 با دقت بیشتری نسبت به سایر الگوریتم ها این کار را انجام می دهد.

۴-۲ پیشنهادها

امروزه با گسترش علم یادگیری ماشین همچنان می توان امیدوار بود که با بهبود هر کدام از الگوریتم های موجود بتوان به دقت بسیار بیشتری نسبت به آنچه تا کنون بدست آمده برسیم. امید است با مطالعه ی بیشتر بر روی الگوریتم های یادگیری ماشین و ترکیب یا بهبود روش های موجود به دقت بالاتری در طبقه بندی ترافیک و درنتیجه امنیت و کیفیت بالاتری برای شبکه ی اینترنت برسیم.

منابع و مراجع

[1] Internet assigned numbers authority (iana). https://www.iana.org/assignments/protocol-numbers/protocol-numbers.xhtml. Accessed: 2021-05-01.

[2] Jamuna, A et al. Efficient flow based network traffic classification using machine learning. 2013.

[3] Shafiq, Muhammad, Yu, Xiangzhan, Laghari, Asif Ali, Yao, Lu, Karn, Nabin Kumar, and Abdessamia, Foudil. Network traffic classification techniques and comparative analysis using machine learning algorithms. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 2451–2455. IEEE, 2016.

[۴] خویی، زهره امینی. طبقه بندی ترافیک شبکه با استفاده از الگوریتم جنگل تصادفی بهبودیافته. ۱۳۹۶.