



دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)  
دانشکده مهندسی کامپیوتر

گزارش درس روش پژوهش و ارائه

طبقه بندی ترافیک شبکه  
با استفاده از الگوریتم‌های یادگیری ماشین

نگارش

محمد مهدی هجرتی

استاد راهنما

دکتر رضا صفا بخش

اردیبهشت ۱۴۰۰





دانشگاه صنعتی امیرکبیر  
(پلی تکنیک تهران)  
دانشکده مهندسی کامپیوتر

گزارش درس روش پژوهش و ارائه

طبقه بندی ترافیک شبکه  
با استفاده از الگوریتم‌های یادگیری ماشین

نگارش

محمد مهدی هجرتی

استاد راهنما

دکتر رضا صفا بخش

اردیبهشت ۱۴۰۰

پاس‌گزاری

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

محمد مهدی جبرتی

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 داده‌های 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

۱	مقدمه	۱
۲	انواع روش‌های طبقه‌بندی ترافیک شبکه	۴
۱-۲	روش مبتنی بر درگاه	۵
۲-۲	روش مبتنی پیلود	۶
۳-۲	روش مبتنی بر رفتار میزبان	۶
۴-۲	روش مبتنی بر یادگیری ماشین	۷
۱-۴-۲	یادگیری نظارت شده	۷
۲-۴-۲	یادگیری بدون نظارت	۷
۵-۲	خلاصه	۸
۳	مدل سازی روش یادگیری ماشین	۹
۱-۳	جمع آوری داده	۱۰
۲-۳	استخراج ویژگی‌ها	۱۱
۳-۳	یادگیری نمونه	۱۱
۴-۳	پیاده سازی الگوریتم	۱۱
۵-۳	بررسی و تحلیل نتایج	۱۱
۶-۳	خلاصه	۱۲
۴	نتیجه گیری و پیشنهادها	۱۳
۱-۴	نتیجه گیری	۱۴
۲-۴	پیشنهادها	۱۴
۱۵	منابع و مراجع	۱۵

## فهرست اشکال

صفحه

شکل

۱-۳ مدل پنج مرحله ای طبقه بندی ترافیک ..... ۱۰

## فهرست جداول

صفحه

جدول

- |     |  |    |
|-----|--|----|
| ۱-۲ | شماره درگاه‌های اختصاص داده شده به برخی از برنامه‌های پرکاربرد | ۵  |
| ۲-۲ | نمونه امضاهای موجود در بسته های برخی از برنامه های پرکاربرد    | ۶  |
| ۱-۳ | دقت طبقه بندی الگوریتم های مختلف یادگیری ماشین                 | ۱۲ |





# فصل اول

## مقدمه

## مقدمه

امروزه طبقه بندی ترافیک شبکه به یک موضوع مهم در حوزه ی کامپیوتر تبدیل شده است. برای ارائه دهندگان خدمات اینترنت<sup>۱</sup>، آگاهی از برنامه های اجرا شده در شبکه یک امر حیاتی می باشد. طبقه بندی ترافیک شبکه اولین مرحله برای تجزیه و تحلیل و شناسایی انواع مختلف برنامه های شبکه است. با این روش ارائه دهندگان خدمات اینترنتی یا اپراتورهای شبکه می توانند عملکرد کلی یک شبکه را مدیریت کنند.

روش های اولیه ی طبقه بندی ترافیک اینترنت مبتنی بر بازرسی بسته<sup>۲</sup> های جریان بودند. روش مبتنی بر شماره درگاه<sup>۳</sup>، شماره درگاه در سرآیند<sup>۴</sup>، 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<sup>۵</sup>، 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 در نتیجه، 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 معمولاً این طور نیست.

امروزه متداول ترین تکنیک مورد استفاده، یادگیری ماشین<sup>۶</sup> است، که توسط بسیاری از محققان استفاده می شود و باعث بدست آمدن نتایج به مراتب دقیق تری از روش های پیشین شده است. تکنیک های یادگیری ماشین، با استفاده از مجموعه ویژگی های آماری جریان به طور خودکار الگوهای ساختاری موجود در انتقال داده های 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 مقایسه ی روش های پیشنهادی موجود

<sup>1</sup> Internet service providers

<sup>2</sup> packet

<sup>3</sup> port

<sup>4</sup> header

<sup>5</sup> payload

<sup>6</sup> 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 می‌کنیم.

## ۱-۲ 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 درگاه‌های 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. مثلاً برنامه‌های وب از پورت ۸۰ استفاده می‌کنند. [۱]

جدول ۱-۲: 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 برنامه‌های 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<sup>۱</sup> استفاده شده است را می‌بینیم.

جدول ۲-۲: نمونه امضاها موجود در بسته‌های برخی از برنامه‌های 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 نمی‌توان آن را بر روی بسته‌های رمزگذاری شده اعمال کرد. به علاوه تجزیه و تحلیل مستقیم داده‌ها باعث نقض حریم خصوصی کاربران می‌شود. همچنین این روش چون محتویات تمام بسته‌ها را بررسی می‌کند، نیازمند سیستم پردازشی به مراتب قوی تری نسبت به روش‌های دیگر است.

## ۳-۲ روش مبتنی بر رفتار میزبان

ایده‌ی اصلی این روش این است که برنامه‌های مختلف الگوهای اتصال متفاوتی دارند. روش مبتنی بر رفتار میزبان، ارتباطات میان میزبان‌های خاص در یک شبکه را شناسایی می‌کند و الگوی ارتباط یک میزبان خاص با الگوی رفتار فعالیت‌های متفاوت مقایسه می‌شود. این روش پیلود بسته را برای طبقه‌بندی ترافیک استفاده نمی‌کند. پس می‌تواند بسته‌ها با محتوای رمزگذاری شده را نیز شناسایی کند. اما اشکال این روش این است که بیشتر تکنیک‌های مبتنی بر رفتار میزبان فرض می‌کند که میزبان

<sup>1</sup>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 برنامه‌های زیادی به طور همزمان استفاده می‌کنند. مشکل دیگری که در این روش بوجود میاد این است که برخی الگوهای رفتاری برنامه را نمی‌توان به آسانی کشف کرد. به مقدار حافظه و جریان‌های زیادی برای همه میزبان‌ها نیاز دارد تا بتواند الگوی اتصال را کشف کند. [۴]

## ۴-۲ روش مبتنی بر یادگیری ماشین

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

### ۱-۴-۲ یادگیری نظارت شده

روش‌های یادگیری با نظارت، مبتنی بر دانش از پیش تعریف شده هستند. این الگوریتم‌ها در مرحله آموزش، نمونه‌های از پیش طبقه‌بندی شده (متشکل از ویژگی‌ها و برچسب مرتبط با آنها) را به عنوان ورودی می‌گیرند و قوانین طبقه‌بندی ایجاد می‌شود. در مرحله طبقه‌بندی تلاش می‌کنند تا برچسب نمونه‌های بدون برچسب را پیش‌بینی کنند.

### ۲-۴-۲ یادگیری بدون نظارت

روش‌های یادگیری بدون نظارت، در مرحله یادگیری خود به هیچ دانش از قبل تعیین شده‌ای نیاز ندارند. این روش گروه‌های طبیعی را در داده‌ها کشف می‌کنند و روی کشف الگوهای موجود در داده‌ها تمرکز دارند. نمونه‌ها را بر اساس میزان شباهت ویژگی‌های آنها که توسط یک رویکرد اندازه‌گیری فاصله تعریف می‌شود. بنابر این در خروجی نمی‌توانیم نوع داده را به طور دقیق مشخص کنیم. صرفاً داده‌های مشابه با هم در یک دسته قرار می‌گیرند.



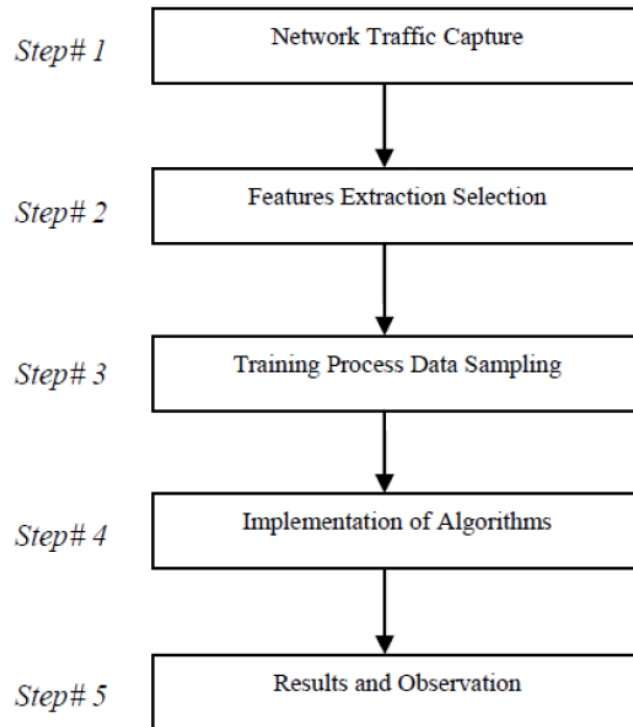
## ۵-۲ خلاصه

در این فصل روش‌های کلی موجود برای طبقه‌بندی 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 برنامه‌های نظیر به نظیر دقت بالایی ندارد. روش مبتنی بر پیلود نیازمند سیستم پردازشی قوی می‌باشد و در شناسایی بسته‌های رمزگاری شده عاجز است. روش مبتنی بر رفتار میزبان صرفاً برای زمان‌هایی که میزبان تنها از یک برنامه در لحظه استفاده می‌کند نتایج دقیقی تولید می‌کند. سپس روش‌های مبتنی بر یادگیری ماشین و انواع آن معرفی شد. در ادامه روش مدل‌سازی و استفاده از یادگیری ماشین در طبقه‌بندی ترافیک شبکه توضیح داده می‌شود. و سپس چند مورد از الگوریتم‌های معروف یادگیری ماشین و نیز پژوهش‌های 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 می شود. شکل زیر خلاصه ای از این مراحل را نشان می دهد.



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

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

### ۱-۳ جمع آوری داده

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

<sup>۱</sup>packet capture

<sup>۲</sup>Wireshark

<sup>۳</sup>tcpdump

## ۲-۳ استخراج ویژگی ها

این مرحله که در پژوهش های مختلف غالبا با ابزارهای نت میت<sup>۴</sup> و پرل اسکریپت<sup>۵</sup> انجام می گیرد، ویژگی های خاصی از بسته ها استخراج می شود و با کمک آن ها دسته بندی یادگیری ماشین<sup>۶</sup> را آموزش می دهند. از جمله ویژگی های قابل استخراج می توان تعداد بسته ها، طول هر بسته، درگاه و پروتکل<sup>۷</sup> 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 می شود تا بتوان به کمک آنها، بسته های ناشناخته را طبقه بندی کرد.

## ۴-۳ 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<sup>۸</sup> این کار را انجام می دهند.

## ۵-۳ بررسی و تحلیل نتایج

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

<sup>۴</sup>Netmate

<sup>۵</sup>Perl script

<sup>۶</sup>machine learning classifier

<sup>۷</sup>protocol

<sup>۸</sup>Weka classification simulation tools

<sup>۹</sup>RBF decision tree

<sup>۱۰</sup>support vector machine (SVM)

همسایه<sup>۱۱</sup>، نیویز<sup>۱۲</sup>، و شبکه بیز<sup>۱۳</sup> نتایج زیر حاصل شده است. [۲، ۳]

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 تقریباً ۸۰ درصدی را نشان می دهد. پس از آن برای الگوریتم های شبکه بیز، ماشین بردار پشتیبان و نیویز 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 می کند.

<sup>11</sup>nearest neighbor

<sup>12</sup>naive bayes

<sup>13</sup>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 می شده است، مورد بررسی قرار گرفت و در مورد مزایا و معایب هر کدام 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 در نتیجه 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.
- [4] خویی، زهره امینی. طبقه بندی ترافیک شبکه با استفاده از الگوریتم جنگل تصادفی بهبودیافته. ۱۳۹۶.