چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) ایران، تهران – ۲۰ و ۲۱ اسفندماه ۱۳۸۷

طراحی و پیاده سازی یک کنترلرفازی برای دسته بندی جریانهای ترافیکی شبکه

علی اصغر یاری فرد
دانشگاه آزاد اسلامی واحد قاینات
a_yarifard@yahoo.com

محمد حسین یغمایی مقدم
دانشگاه فردوسی مشهد
h.yaghmaee@yahoo.com

چکیده: شناسایی و دسته بندی دقیق جریانهای ترافیکی شبکه بر اساس پروتکل تولید کننده آنها، یکی از المانهای اصلی مدیریت شبکه محسوب می شود. تاکنون تکنیک های مختلفی در این زمینه مانند شماره پورت و بررسی بخش داده ای بسته ها ارائه شده است. امروزه بسیاری از برنامه ها (مانند P2P) با استفاده از شماره پورتهای پویا و تکنیک های رمزنگاری سبب ناکارآمدی تکنیکهای فوق شده اند. تکنیک دیگر در این زمینه، استفاده ازاطلاعات آماری لایه انتقال جریانهای عبوری می باشد. در این مقاله یک کنترلر فازی برای دسته بندی جریانهای ترافیکی پیشنهاد شده است. در این کار اطلاعات آماری لایه انتقال جریانهای عبوری، بعنوان پارامترهای دسته بندی استفاده شده است. نتایج آزمایشات نشان می دهند که مکانیزم پیشنهادی بطور متوسط دارای کارایی کلی ۹۴٪ بوده و ۹۸.۷۷٪ جریانها را شناسایی کرده است. در حالی که کارایی کلی الگوریتم های K-Means و DBSCAN به ترتیب ۸۴٪ و ۷۵.۶٪ می باشد.

واژه های کلیدی: شناسایی جریانهای ترافیکی شبکه، سیستم های فازی

۱- مقدمه

با گسترش روز افزون تعداد پروتکلهای لایه کاربرد و انواع کاربران نهایی، مدیریت کارای منابع شبکه یکی از مشکلات اصلی مدیریت شبکه های بزرگ می باشد. مدیریت مناسب یک شبکه، نیازمند داشتن دید جامعی از وضعیت فعلی شبکه و نحوه استفاده جریانهای عبوری از منابع شبکه می باشد. مکانیزمهای دسته بندی جریانهای ترافیکی شبکه، ابزارهای مناسبی هستند که در زمینه های مختلف مدیریت شبکه از جمله مدیریت منابع (مانند تخصیص و کنترل منابع) و بهبود امنیت سیستم های تشخیص حملات شبکه[1] می توان از آن استفاده کرد.

برای دسته بندی جریانهای ترافیکی شبکه بر اساس پروتکل تولید کننده آنها، تکنیک های مختلفی مانند شماره پورت (port-based) و بررسی بخش داده ای بسته های عبوری(payload-based)، ارائه شده است. در روش اول از شماره پورت مبدأ و مقصد بسته ها در لایه انتقال[۱] و در روش دوم از بررسی بخش داده ای بسته ها [۲] جهت شناسایی و دسته بندی جریانهای ترافیکی عبوری استفاده می شود. تحقیقات اخیر نشان می دهد که عواملی مانند استفاده از تکنیک های رمزنگاری، اجرای سرویسهای استاندارد بر روی پورتهای غیر استاندارد، و گسترش روزافزون برنامه های که از شماره پورتهای پویا بهره می گیرند، سبب ناکارآمدی این تکنیکها شده است[۳].

استفاده از اطلاعات آماری جریانها در لایه انتقال، تکنیک دسته بندی دیگری است که برای غلبه بر محدودیتهای روشهای قبلی ارائه شده است[۴، ۵]. تکنیک مذکور بر این اصل استوار است که برنامه های کاربردی مختلف هنگام برقراری ارتباط، دارای الگوهای رفتاری مختلفی هستند. تحقیقات انجام شده در این زمینه نشان می دهد که با استفاده از اطلاعات آماری لایه انتقال مانند تعداد کل بسته ها، تعداد کل بایتهای ارسالی، متوسط طول بسته ها و ... می توان رفتار برنامه های کاربردی مختلف را شناسایی کرد.

در این مقاله یک کنترلر فازی برای دسته بندی جریانهای ترافیکی شبکه پیشنهاد شده است. در مکانیزم پیشنهادی از اطلاعات آماری جریانها ترافیکی در لایه انتقال بعنوان فاکتورهای دسته بندی بهره گرفته شده است. عملکرد روش پیشنهادی از دیدگاههای مختلفی مورد بررسی قرار گرفته است. نتایج آزمایشات نشان می دهد که روش پیشنهادی در حدود ۹۸.۷۷٪ جریانها را شناسایی کرده و کارایی کلی روش پیشنهادی بطور متوسط حدود ۹۴٪ می باشد. در حالی که کارایی کلی الگوریتم های K-Means و DBSCAN به ترتیب ۸۴٪ و ۷۵.۶٪ می باشد.

۲- کارهای مرتبط

تاکنون تحقیقات مختلفی در زمینه دسته بندی جریانهای ترافیکی شبکه با استفاده از اطلاعات آماری لایه انتقال انجام شده که در ادامه بطور مختصر چندین کار جدید در این زمینه بررسی می شود.

McGregor جریانهای ترافیکی را با بکارگیری تکنیک های کلاسترینگ و استفاده از اطلاعات آماری جریانها در لایه انتقال دسته بندی کرد. در این کار کارایی و دقت روش استفاده شده مورد بررسی

1- Network Intrusion Detection System

قرار نگرفته و تنها به بررسی انتخاب خصوصیات آماری مورد نیاز و تاثیر آن بر دسته بندی پرداخته شده است[۵].

در [۶]، Roughen با استفاده از تنها دو خصوصیت آماری طول مدت ارتباط و متوسط طول بسته ها، جریانهای ترافیکی را در چهار دسته مختلف دسته بندی کرد. در این کار برای دسته بندی جریانها از الگوریتم های بررسی نزدیکترین همسایه و جداکننده خطی استفاده شده است. ایراد اصلی روش استفاده شده آن است که با استفاده از تنها دو خصوصیت آماری نمی توان جریانهای ترافیکی برنامه های کاربردی مختلف را از هم تفکیک کرد.

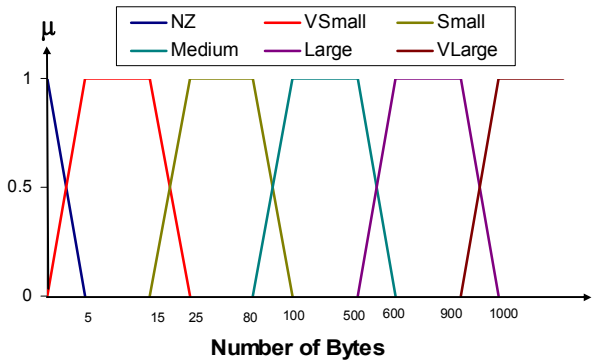
Erman در [۷]، به ارزیابی سه الگوریتم کلاسترینگ AutoClass، K-Means و DBSCAN در زمینه شناسایی جریانهای ترافیکی پرداخته است. در این کار از اطلاعات آماری لایه انتقال جریانها بعنوان پارامترهای دسته بندی استفاده شده است. همچنین عملکرد مکانیزمهای بیان شده مورد ارزیابی و مقایسه قرار گرفته است.

۳- مکانیزم دسته بندی پیشنهادی

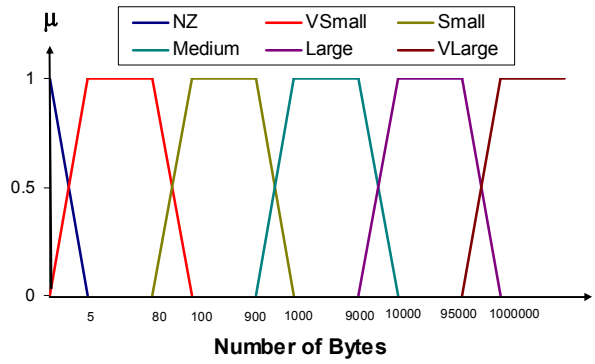
در این بخش ما مکانیزم دسته بندی پیشنهادی را که بر اساس منطق فازی طراحی شده را شرح می دهیم. مکانیزم دسته بندی مذکور که FC[2] نامگذاری شده، در واقع یک کنترلر فازی است که از چهار بخش اصلی فازی ساز، قوانین فازی، موتور استنتاج فازی و غیرفازی ساز تشکیل شده است.

سیستم فازی مورد نظر دارای شش پارامتر ورودی است که از سه متغیر زبانی مجزا بهره می گیرد. متغیرهای زبانی پارامترهای ورودی سیستم شامل تعداد کل بایتها، تعداد کل بسته ها و متوسط طول بسته ها در هر طرف جریانها می باشد. همچنین در این سیستم از یک متغیر زبانی برای توصیف تنها پارامتر خروجی سیستم استفاده می کنیم. توابع عضویت متغیرهای زبانی پارامترهای ورودی و خروجی در شکل ۱، ۲، ۳و ۴ نشان داده شده است.

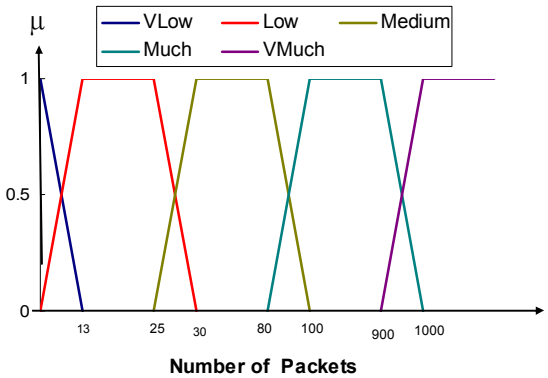
مهمترین بخش طراحی سیستم فازی فوق، طراحی و تولید قوانین مناسب و دقیق برای شناسایی کلاسهای ترافیکی مختلف می باشد. در این کار، برای تولید قوانین مورد نظر از داده های آموزشی استفاده شده است. قالب قوانین فازی طراحی شده بصورت IF-THEN می باشد. بخش IF قوانین یک گزاره فازی است که از ترکیب مقادیرمتغیرهای زبانی پارامترهای ورودی با عملگرفازی and تشکیل شده است. بخش THEN نیز نتیجه قانون اعمال شده را نشان می دهد که در سیستم ما بصورت مجموعه فازی می باشد. در بخش موتور استنتاج، از عملگر ضرب برای استنتاج از قوانین فازی و تفسیر عملگر فازی and استفاده شده است. برای ادغام نتایج حاصل از قوانین فازی نیز از عملگر اجتماع استفاده شده است. برای ساده سازی محاسبات موتور استنتاج از مجموعه های فازی Singleton در بخش فازی ساز بهره گرفته شده است. با توجه به اینکه در طراحی سیستم مذکور از توابع عضویت Singleton برای متغیر زبانی پارامتر خروجی استفاده شده، از تکنیک ماکزیمم گیری[3] برای نگاشت تنها پارامتر خروجی سیستم پیشنهادی استفاده شده است.



شکل۱- توابع عضویت متغیر زبانی تعداد بایتهای ارسالی



شکل۲- توابع عضویت متغیر زبانی میانگین طول بسته



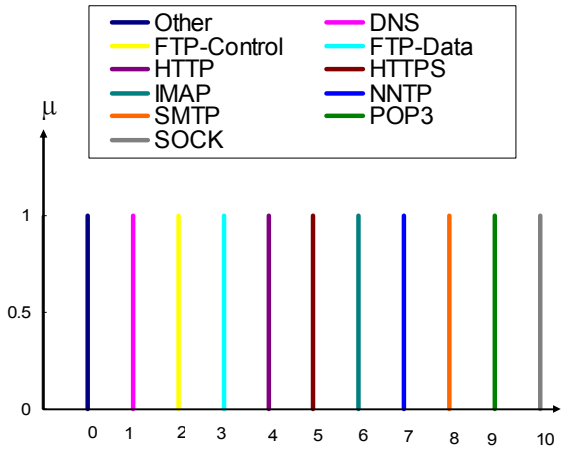
شکل۳- توابع عضویت متغیر زبانی تعداد بسته های ارسالی

۴- داده های آموزشی

برای شناسایی جریانهای ترافیکی بایستی مکانیزمهای دسته بندی

2 - Fuzzy Classifier

3 - Maximum Label



شکل ۴- تابع عضویت پارامتر خروجی سیستم فازی

مورد استفاده با نمونه های مناسب از کلاسهای ترافیکی مختلف آموزش داده شوند. بنابراین میزان دقت مکانیزم دسته بندی به داده های آموزشی وابسته است. داده های آموزشی مورد استفاده برای آموزش مکانیزمهای دسته بندی بایستی از داده های ترافیکی عملی نمونه برداری شوند. در این تحقیق برای جمع آوری داده های آموزشی، از داده های ترافیکی عبوری از خط اینترنت دانشگاه اوکلند(که به Auckland IV[۴] مشهور است) استفاده شده است. داده های ترافیکی مذکور تنها شامل سرآیند TCP/IP یا UDP/IP بسته های مربوط به جریانهای ترافیکی عبوری مختلف بوده که در بازه زمانی۲۰۰۱/۳/۱۶ تا ۲۰۰۱/۳/۱۹ نمونه برداری شده است. در این کار، از مقادیر فیلدهای سرآیند بسته ها مانند آدرس مبدأ، آدرس مقصد، شماره پورت مبدأ، شماره پورت مقصد و شماره پروتکل مورد استفاده در لایه انتقال، برای شناسایی بسته های جریانهای ترافیکی مختلف استفاده می شود. با استفاده از ابزار tcptrace [۸] ، تعداد ۴،۲۵۰،۰۰۰ جریان ترافیکی مختلف شناسایی و اطلاعات آماری مربوط به آنها جمع آوری شد. جریانهای شناسایی شده بر اساس شماره پورتهای استاندارد به کلاسهای ترافیکی مختلفی تقسیم شدند. دلیل استفاده از روش دسته بندی جریانها بر اساس شماره پورت آن است که این داده های ترافیکی قبل از رواج برنامه های است که از تکنیک شماره پورتهای پویا استفاده می کنند. اطلاعات جریانهای جمع آوری شده بطور خلاصه در جدول ۱ نشان داده شده است.

۵- نتایج آزمایشات

برای ارزیابی عملکرد مکانیزمهای دسته بندی از فاکتورهای مختلفی استفاده می کنند. برخی از این فاکتورها عبارتند از: کارایی کلی مکانیزم دسته بندی در شناسایی صحیح جریانهای ترافیکی، میزان دقت مکانیزم در شناسایی جریانهای ترافیکی هر کلاس ترافیکی. در این

جدول۱- اطلاعات جریانهای جمع آوری شده از داده های ترافیکی

نام پروتکل	تعداد جریانها	تعداد بایتها	درصد بایتها
HTTP	۳،۸۵۰،۰۳۹	۳۵،۵۴۴،۴۲۲،۳۲۵	٪۷۶.۳۴
SMTP	۱۲۰،۰۲۷	۳،۰۸۱،۹۰۵،۸۰۷	٪۶.۲۷
HTTPS	۱۶۴،۶۵۳	۱،۰۶۰،۵۲۹،۰۰۰	٪۲.۱۵
FTP-Dat	۷،۸۲۲	۹۳۸،۴۵۳،۱۱۳	٪۱.۹۱
NNTP	۱۴،۶۳۶	۵۵۲،۲۴۲،۲۱۹	٪۱.۱۲
SOCK	۱۲۰،۰۲۷	۲۴۹،۲۷۵،۴۹۱	٪۰.۵۴
POP3	۵۳،۹۵۰	۱۶۹،۶۴۴،۴۶۷	٪۰.۳۴
DNS	۱۳،۶۵۹	۷۳۳،۱۲۶،۶۹۸	٪۰.۱۴
IMAP	۳،۴۲۴	۳۹،۳۰۱،۰۵۲	٪۰.۰۷
FTP-Con	۵،۱۴۸	۸،۸۸۵،۳۵۹	٪ ۰.۰۱
OTHER	۲۳۱،۰۱۷	۵،۴۳۰،۲۲۲،۲۰۸	٪۱۱.۰۴

بخش ما عملکرد مکانیزم دسته بندی پیشنهادی FC را با مکانیزمهای K-Means و DBSCAN مورد ارزیابی و بررسی قرار داده ایم. به همین منظور آزمایشاتی با تعداد جریانهای ترافیکی مختلف صورت گرفته است. در این آزمایشات، مجموعه جریانهای ترافیکی آزمایشی، به دسته های با ۲۵۰،۰۰۰ نمونه جریان ترافیکی تقسیم شدند. آزمایشات با مقدار اولیه ۲۵۰،۰۰۰ نمونه شروع و در هر تکرار آزمایش، تعداد ۲۵۰،۰۰۰ نمونه دیگر به داده های آزمایشی افزوده شده است.

۵-۱ ارزیابی کارایی کلی

برای ارزیابی کارایی کلی مکانیزم پیشنهادی FC از پارامتر دقت کلی[۵] استفاده شده است. معیار دقت کلی برای یک مکانیزم دسته بندی طبق فرمول زیر تعریف می شود:

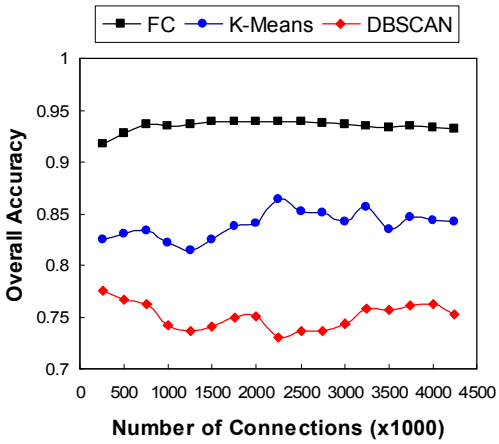
$$overall\ accuracy = \frac{\sum TP \quad for \quad all \quad traffic \quad class}{total \quad number \quad of \quad connections} \qquad (1)$$

که در آن TP بیانگر تعداد کل جریانهای ترافیکی است که توسط مکانیزم دسته بندی بدرستی شناسایی شده است.

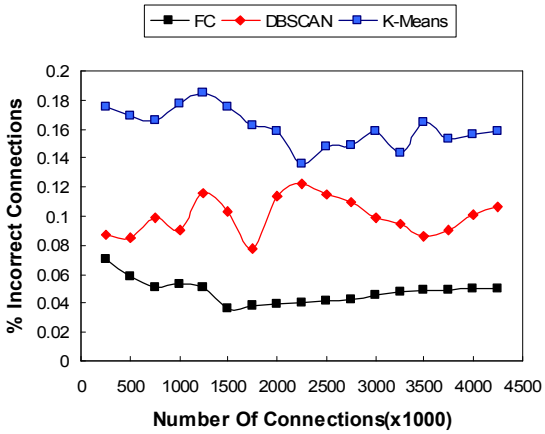
نتایج آزمایشات انجام شده بیانگر آن است که دقت کلی شناسایی جریانها در مکانیزم FC به ترتیب بطور متوسط حدود ۹۴٪ می باشد. در حالی که دقت کلی شناسایی جریانها در مکانیزم K-Means بطور متوسط در حدود ۸۴٪ ، و در مکانیزم DBSCAN برابر ۷۵.۶٪ و ۷۰.۵۱٪ می باشد(شکل ۵). همچنین شکل ۶ نشان می دهد که مکانیزمهای FC ، K-Means و DBSCAN به ترتیب بطور متوسط حدود ۴.۷۸٪ ، ۱۶.۱٪ و ۹.۸۱٪ جریانها را نادرست شناسایی و دسته بندی کرده اند.

4 - ftp://pma.nlanr.net/traces/long/auck/4/

5- Overall Accuracy



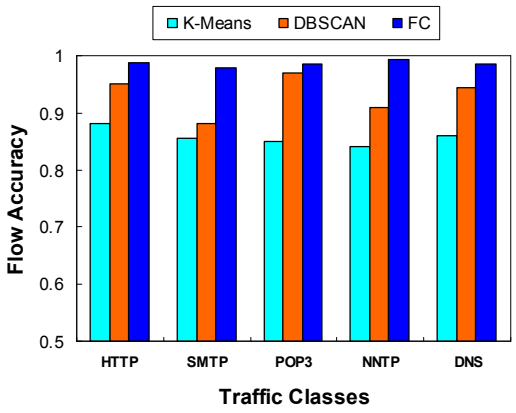
شکل۵- دقت کلی مکانیزمهای دسته بندی



شکل۶- نمودار جریانهای شناسایی شده نادرست

۲-۵ ارزیابی دقت هر کلاس ترافیکی

فاکتور کارایی کلی که در بخش قبلی مورد بررسی قرار گرفت، در واقع معیاری برای ارزیابی دقت کلی یک مکانیزم دسته بندی می باشد. این معیار دید واضح و مشخصی از توانایی مکانیزم در شناسایی صحیح جریانهای مربوط به کلاسهای ترافیکی مختلف ارائه نمی کند. در این بخش برای ارزیابی بیشتر توانایی مکانیزمهای دسته بندی فوق در شناسایی جریانهای ترافیکی مربوط به هر کلاس ترافیکی، بررسیهای بر روی نتایج آزمایشات بخش قبلی صورت گرفت. این بررسیها بدون لحاظ جریانهای است که توسط مکانیزمهای دسته بندی بصورت UNKNOWN شناخته شده و تنها جریانهای که در کلاسهای ترافیکی مختلف قرار گرفته اند مورد توجه بوده است. دقت هرکلاس ترافیکی بصورت نسبت کل جریانهای ترافیکی که بدرستی در هر کلاس ترافیکی قرار گرفته به کل جریانهای موجود در هر کلاس ترافیکی تعریف می شود. نتایج بررسیها در شکل ۷ نشان داده شده است. نتایج نشان می دهند که مکانیزمهای دسته بندی FC، K-Means و DBSCAN جریانهای ترافیکی مربوط به کلاسهای ترافیکی مختلف را به ترتیب بطور متوسط با دقتی در حدود ۹۷٪ ، ۹۵٪ و ۸۵٪ شناسایی می کنند.



شکل ۷- دقت شناسایی جریانهای کلاسهای ترافیکی

۶. نتیجه گیری

در این مقاله برای شناسایی و دسته بندی جریانهای ترافیکی شبکه، یک کنترلر فازی طراحی و پیاده سازی شده است. در این کار، مکانیزم پیشنهادی با مکانیزمهای دسته بندی دیگر مانند K-Means و DBSCAN مقایسه شده است. آزمایشات انجام شده نشان می دهد که کارایی کلی مکانیزم دسته بندی پیشنهادی در مقایسه با مکانیزمهای دیگر بالاتر می باشد. همچنین مکانیزم مذکور قابلیت شناسایی جریانهای ترافیکی نامشخص را نیز دارد.

مراجع

[1] Moore, D., Keys, K., Koga, R., Lagache, E., Claffy Kimberly C., "The CoralReef Software Suite as a Tool for Systems and Network Administrators", *In LISA '01: Proceeding of the 15th USENIX conference on Systems Administration*, pp. 133-144, San Diego, CA, USA, December 2001.

[2] Paxson, V., "Bro: A System for Detecting Network Intruders in Real-Time", *In Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, January 1998.

[3] Sen, S., Spatscheck, O., Wang, D., "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures", *In WWW2005*, New York, USA, May 17-22, 2004.

[4] Moore, Andrew W, and Zuev, D, "Internet Traffic Classification Using Bayesian Analysis Techniques", *In SIGMETRIC'05*, Banff, Canada, June 2005.

[5] McGregor, A., Hall, M., Lorier, P., Brunskill, J., "Flow Clustering Using Machine Learning Techniques", *In PAM'04*, Antibes Juan-les-Pins, France, April 19-20, 2004.

[6] Roughan, M., Sen, S., Spatscheck, O., Duffield, N., "Class-of-Service Mapping for QoS: A Statistical Signature-based Approach to IP Traffic Classification", *In IMC'04*, Taormina, Italy, October , 2004.

[7] Erman, J., Mahanti, A., Arlitt, M., "Traffic Classification Using Clustering Algorithms", *In SIGCOMM'06 MineNet Workshop*, Pisa, Italy, September 2006.

[8] http://www.tcptrace.org/index.html