网络流量分类测度的相关性分析[1]

胡晓艳* 龚俭

（东南大学计算机科学与工程学院；江苏省计算机网络技术重点实验室 南京 210096）

摘要：

理论上说，网络流量分类测度越多，就能从更多的角度刻画流量以便更精确地分类流量。但过多的测度使它们之间的关系复杂化，且其测量及处理耗费大量资源。通过对流量分类测度的数据分析后发现，不少测度之间存在相关性，因此本文提出使用变量聚类，基于测度之间的相关性，将 181 个测度分成了 32 类，即将具有强相关性测度自动聚类，理清了测度之间的关系，并发现了一些潜在的事实，同时找到了每类测度的代表性测度;然后使用回归分析建立每类的代表性测度与该类中其他测度之间的量化关系，在确保一定精度的情况下实现测度之间的互推，进而减少部分测度的测量开销。

关键字：网络流量分类测度 相关性分析 变量聚类 回归分析

中图分类号：TP393.0 **文献标识码**：A

Correlation Analysis for Discriminators in Flow Classification

XIAOYAN HU, JIAN GONG

School of Computer Science & Engineering, Southeast University; Key Laboratory of Computer Network Technology in Jiangsu, Nanjing 210096, China

Abstract:

In theory, the more discriminators for flow classification, the more aspects of flows they would characterize so that the accuracy of classification process would be higher. However, it is difficult to figure out the relationships among so many discriminators and it is expensive in measuring and processing these discriminators. Fortunately, it is found that some discriminators are highly correlated. And thus, in this paper, variable clustering is introduced to clarify relationships by clustering 181 discriminators into 32 clusters based on the correlation between them and selecting a representative discriminator for each cluster. Then regression analysis is used to establish quantified relationship between a cluster representative and other discriminator in the cluster so as to predict the discriminator from the representative and eventually reduce measuring cost.

Keywords: discriminators in flow classification, correlation analysis, variable clustering, regression analysis

1. 背景

精确的网络流量分类对于一些网络活动是必备的，如网络安全监测、流量计费、网络规划等。网络流量的精确分类需要大量测度刻画网络流量的各方面以区分不同类别的流量，因而 Andrew Moore 提出了多达 248 个流量分类测度来刻画每个流量。网络流量分类测度是符合特定流规范和超时约束的报文流的独特特征或描述该流行为的参数，因其具有将网络流量分类的特性，故又被称为鉴别器（discriminator）[1]。

网络流量分类测度的选取对于网络流量分类的合理性及精确性至关重要。一般来说，流量分类测度越多，就能从越多的角度来精确地刻画各类流量的特征进而提升分类的精确性。但与此同时，网络流量分类进程的复杂性及所消耗的资源是需要考虑的问题。大量的测度，尤其是某些测度间不正交时，会使分类问题复杂化甚至影响分类的正确性。本文的目标是通过剖析测度之间的关系，简化网络流量分类进程并减少不必要的开销。

在研究测度之间的关系[2]时，必然有几个问题待以解决：如何发现哪些测度之间存在相关性？具有相关性的两个测度中哪个应该作为自变量？两个相关测度之间的关系是什么样的关系？本文研究的网络流量分类测度集合所包含的元素多达 248 个，测度之间的关系错综复杂，本文提出使用变量聚类和回归分析（VCRA）的组合回答上述三个问题。最终将网络流量分类测度中具有

[1]本文工作由国家科技支撑项目（No.2008BAH37B04）及国家重点基础研究发展 973 计划（No.2009GB320505）资助。作者简介：胡晓艳（1985.7-），女，博士生，网络管理专业；龚俭（1957.7-），男，博士，计算机网络教授。

*通讯作者 E-Mail ：xhbreezehu@gmail.com

强相关性的测度组合成类，并为每类测度找到了代表性测度；且以每类测度的代表性测度为自变量，建立了其与该类中其他测度（因变量）之间的量化关系。测度之间的关系进一步明确了，并且测度间量化关系的建立可以减少部分测度的测量开销。

2. VCRA 算法

VCRA 算法流程如图 1 所示。首先将网络流的分类测度的数据输入 SAS 系统中，在变量聚类（在 SAS 中用 VARCLUS 过程实现）模块，得到了网络流量分类测度的相关系数矩阵，该矩阵被用于将具有强相关性的测度分成一类，并找到每类测度中的代表性测度（作为回归分析中的自变量）；然后对代表性测度和类中其它测度的数据使用回归分析，建立两者之间的量化关系。

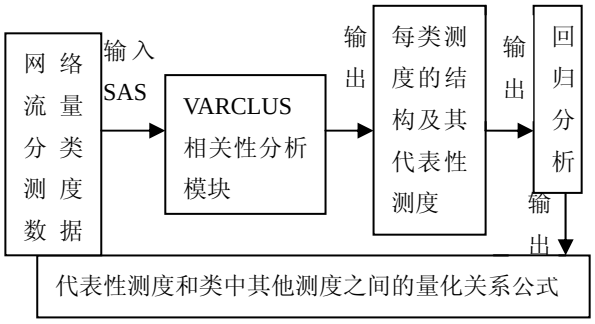

图 1 VCRA 算法流程

2.1 变量聚类

变量聚类是依据网络流量分类测度的相关系数矩阵，对测度（而非数据）进行无监督的聚类，所得到的测度类中的测度之间尽可能地相关，并且与其他类中的测度尽可能地不相关。根据每类测度中的测度与本类测度以及测度与其下一最接近的测度类之间的关系找到该类测度的代表性测度。

在 SAS 中，非监督的变量聚类由 VARCLUS 过程实现，使用斜交多组组成分分析对测度同时进行分割聚类和系统聚类，其开始把所有测度看成一类，之后重复以下步骤：

1）选择一个欲分的类，被选择的类要么是类第一主分量所含信息最少，或其第二特征值在当前所有类中最大；

2)运用所选择类的前两个主分量将该类拆分，对这两个主分量进行正交斜交旋转，然后再把此类中所有的测度归入与这两个主分量相关系数较大者；

3）测度被循环地归入各类以使类成分所占的方差部分到达最大。

当每个测度类成分所占方差百分比或每类第二特征根的上限满足用户指定的要求时，上述过程停止。最后，变量聚类方法规定每类测度中具有最小$1\text{-}R^2ratio$的测度作为该类测度的代表性测度。最佳的类的代表性测度应该与本类测度尽可能地相关，并尽可能与最近的类不相关，即其$1\text{-}R^2ratio$应该趋近于 0。$1\text{-}R^2ratio$如下定义：

$$1-R^2ratio=(1-R^2own)/(1-R^2nearest)$$

其中 R^2own 和 $R^2nearest$ 分别表示该测度与本类、下一最近类之间的相关性[3]。

3. 实验及结果

用于变量聚类的网络流量分类测度需具备以下条件：1）是数值型变量；2）易变的，表明它与网络流量分类相关的。因此 248 个测度中仅 181 个测度用于相关性分析。用于相关性分析的数据是从网络上采集的 65536 条流量，关于这些流量的具体信息参考[4]。

为了缩减篇幅，图 2、3 列出了部分变量聚类的结果。图 2 显示了网络流量分类测度被分成 32 类时的总结性信息。可以看到，例如，第二个类包含 13 个测度，该类测度所解释的总方差为 12.15734，占该类测度总方差的 93.52%；此外，“total variation explained=144.8299”表示所有测度类的主分量所解释的总方差为 144.8299，“Proportion=0.8002”表示该 32 个测度类主成分解释了 80.02%的数据的变化，即意味着网络流量分类测度聚类为 32 类足以刻画原所有测度刻画的数据的大部分变化，则该 32 类测度的代表性测度能够刻画所有 181 个测度的大部分的变化。

图 3 显示了两个测度类的结构。可以看到，所有测度的$1\text{-}R^2ratio$和 $R^2nearest$ 的值都比较小，进一步说明变量聚类的结果比较理想，选择代表性测度是有意义的。根据选择类的代表性测度的规则，第 2、24 两个类的代表性测度分别为测度“actual_data_pkts_ba”（服务器向客户机发送的报文数）和测度“sack_pkts_sent_ba”（服务器到客户机方向具有选择性确认选项的报文数）。此外，通过对每类测度的结构可以发现一些

潜在的事实，如从第 24 类测度所包含的 3 个测度可以发现，网络流量中某一方向的具有选择性确认选项的报文数与另一方向的乱序的报文数目密切相关，这是个可以解释的网络中的事实。

Cluster summary for 32 clusters

Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	13	13	11.34001	0.8723
2	13	13	12.15734	0.9352
3	8	8	7.228801	0.9036
4	6	6	5.726707	0.9545
5	12	12	11.03291	0.9194
6	5	5	3.525023	0.7050
7	7	7	6.246081	0.8923
8	9	9	6.886213	0.7651
9	3	3	3	1.0000
10	6	6	5.05706	0.8428
11	6	6	5.336296	0.8894
12	4	4	2.950441	0.7376
13	3	3	3	1.0000
14	4	4	2.860792	0.7152
15	3	3	1.911927	0.6373
16	3	3	3	1.0000
17	9	9	7.436705	0.8263
18	3	3	2.383513	0.7945
19	3	3	2.863781	0.9546
20	6	6	3.183953	0.5307
21	2	2	2	1.0000
22	5	5	3.198772	0.6398
23	2	2	1.996232	0.9981
24	3	3	1.986267	0.6621
25	3	3	2.519574	0.8399
26	13	13	9.989729	0.7684
27	4	4	1.640728	0.4102
28	2	2	2	1.0000
29	2	2	1.831569	0.9158
30	7	7	6.156165	0.8795
31	3	3	3	1.0000
32	8	8	1.383321	0.1537

Total variation explained=144.8299 Proportion=0.8002

图 2　VARCLUS 过程的聚类总结性信息

Oblique Centroid Component Cluster Analysis

Cluster	Variable	R-squared with Own Cluster	R-squared with Next Closest	1-R**2 Ratio
2	pure_acks_sent_ab	0.9665	0.0893	0.0368
	unique_bytes_sent_ba	0.9593	0.1188	0.0462
	actual_data_pkts_ba	0.9773	0.1208	0.0258
	actual_data_bytes_ba	0.9607	0.1192	0.0446
	rexmt_data_pkts_ba	0.8636	0.1096	0.1532
	rexmt_data_bytes_ba	0.7795	0.1168	0.2496
	ttl_stream_length_ba	0.9593	0.1188	0.0462
	RTT_samples_ba	0.9763	0.1044	0.0265
	RTT_full_sz_smpls_ba	0.9761	0.1002	0.0266
	post_loss_acks_ba	0.9319	0.0750	0.0736
	segs_cum_acked_ba	0.9174	0.1305	0.0950
	duplicate_acks_ba	0.9521	0.0831	0.0523
	triple_dupacks_ba	0.9481	0.0937	0.0572
24	sack_pkts_sent_ba	0.8516	0.0129	0.1503
	max_sack_blks_ba	0.4495	0.0207	0.5622
	outoforder_pkts_ab	0.7188	0.0096	0.2840

图 3　VARCLUS 过程聚类后每个测度的 R-square 值

The REG Procedure Model: MODEL1
Dependent Variable: segs_cum_acked_ba segs_cum_acked_ba

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	32012437427	32012437427	2200744	<.0001
Error	65533	953255470	14546		
Corrected Total	65534	32965692897			

Root MSE	120.60759	R-Square	0.9711
Dependent Mean	9.80917	Adj R-Sq	0.9711
Coeff Var	1229.53908		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1.31779	0.47119	-2.80	0.0052
actual_data_pkts_ba	actual_data_pkts_ba	1	0.51355	0.00034617	1483.49	<.0001

图 4　REG 过程的结果

基于变量聚类所得到的每个测度类的结构及其代表性测度，接下来则可以以每个类的代表性测度为自变量，用回归分析[5]来建立其与该类中其他测度之间的量化关系。本文着重解释回归分析在第二类测度上的分析。图 5 中显示了回归分析的一个例子，即使用 REG 过程建立第二类测度中的 “segs_cum_acked_ba”（从服务器到客户方向的累积确认的报文数目）和类的代表性测度“actual_data_pkts_ba”之间的量化关系的结果。整体模型的 F 统计表明回归分析结果是显著的（F=2200744, p<0.0001），0.9711 的 R-Squre 指标表明测度“actual_data_pkts_ba”能够刻画测度“segs_cum_acked_ba”数据 97.11%的变化，即前者可以预测后者的大部分数据的变化。在“Parameter Estimate”表格中的 p-values（t=-2.80, p=0.0052 and t=1483.49, p<0.0001）表明截距和斜率参数的估计都是比较明显的。因此这两个测度之间的量化关系式如下：

segs_cum_acked_ba=-1.31779+0.51355actual_data_pkts_ba

有了这个量化关系式，这两个测度之间的关系明确了，进而也可以实现这两者之间的互推，减少测度的测量开销。

4. 总结和展望

基于现有的网络流量分类测度数目众多、测度之间关系复杂、耗费大量测量和处理资源的事实，本文提出使用变量聚类将 181 个网络流量分类测度中具有强相关性的网络流量分类测度自动聚类，这样有助于了解哪些测度是相互相关的，并从语义的角度分析测度之间相关的意义；另外，变量聚类还选出每个测度类的代表性测度，为进一步分析类内测度之间具体的关系提供了前提，即类的代表性测度被作为自变量，类中的其他测度作为因变量，然后运用回归分析建立这两者之间的量化关系模型，实现了在确保一定精度的前提下从类的代表性测度推测类中其他测度，最终减少这些被推测的测度的测量开销。

由于测度之间的相关性分析是个复杂的问题，即使知道某些测度相关，它们之间确切的量化关系的寻找仍是个难题，变量聚类、回归分析只是相关性分析工具的一部分，作者下一步的工作是为发现的这些测度之间的相关性建立进一步的定量模型。

参考文献

1. A. W. Moore, D. Zuev, M. Crogan. Discriminators for Use in Flow-based Classification[R]. Technical Report, RR-05-13, Department of Computer Science, Queen Mary, University of London, August, 2005

2. Donald B. Macnaughton: Definition of Relationship Between Variables"[EB/OL]. http: //www.matstat.com /teach/p0045.htm, 2002-01-28.

3. SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*[M]. Cary, North Carolina: SAS Institute Inc., 1999.

4. A. W. Moore and D. Zuev. Internet Traffic Classification Using Bayesian Analysis Techniques[C] //*Proceeding of the 2005 ACM SIGMETRICS Performance Evaluation Review*. New York: ACM, 2005: 50-60

5.Draper, Norman R. *Applied Regression Analysis, 3rd ed*.[M]. New York, Toronto: John Wiley & Sons, 2006.