的影响.本节将平衡训练数据集后的参数选择方法与文献[4]的参数选择方法(未平衡训练数据集)进行比较.由图 7 可见,利用平衡训练数据集后选择的聚类参数识别应用时,将获得较高的检全率,尤其是对样本数较少的 Skype 和 other 类的识别.这主要是由于在平衡后的训练数据集上进行聚类,更有可能生成识别样本数较少应用的簇.因此,平衡训练数据集对 K 均值聚类算法的参数选择是至关重要的.

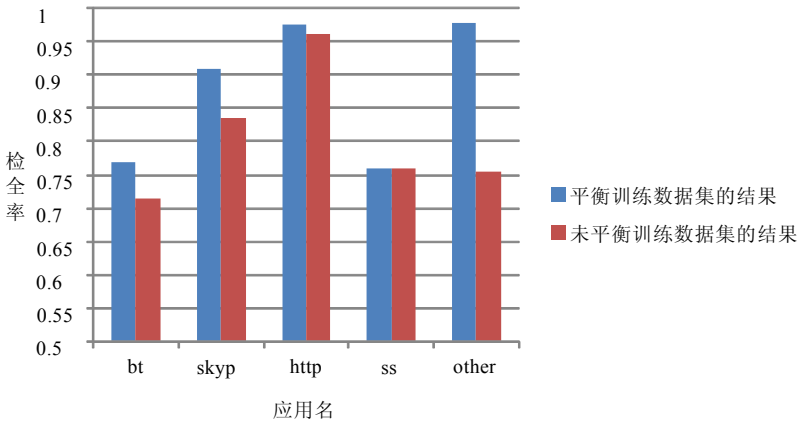


Fig.7 Comparison of TPRs on the balanced training dataset and non-balanced training dataset

图 7 平衡训练数据集与未平衡训练数据集的检全率比较

4.3 TCP单向流观察窗口大小

在确定数据包识别能力和 K 均值聚类算法的参数之后,还需要确定 TCP 观察窗口的大小,以便能够在线识别流量.对于从客户端到服务器方向的流和从服务器到客户端方向的流,本节为 TCFEC 分别确定观察窗口的大小.我们通过改变观察窗口大小使 TCFEC 获得不同的分类准确率,然后选择使分类准确率达到最大的观察窗口大小.

由图 8 和图 9 可见,对于客户端到服务器方向的流,前 3 个数据包可使 TCFEC 达到最高的准确率;而对于服务器到客户端方向的流,前两个数据包获得的准确率最高.当观察窗口变大时,例如从客户端到服务器方向观察 4 个数据包时,分类准确率下降.这很可能是由于第 4 个数据包干扰了 TCFEC 识别流量,导致整体分类准确率下降.因此,对于客户端到服务器方向的流,本节设置观察窗口大小为 3;而对于服务器到客户端方向的流,本节设置观察窗口大小为 2.

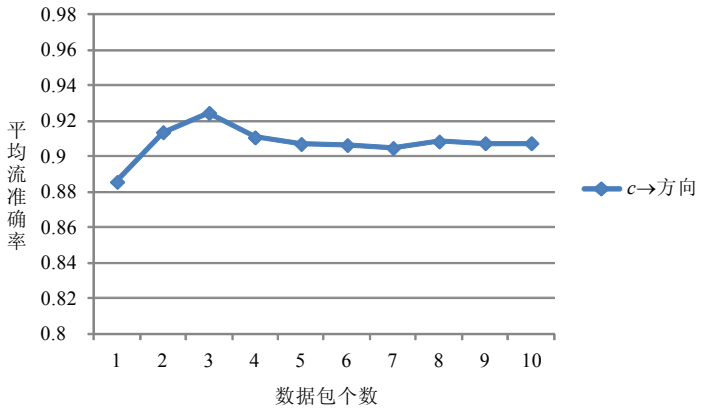


Fig.8 Overall accuracy achieved with the different number of packets in the client-to-server flows

图 8 客户端到服务器方向的流,使用不同的包数目获得的整体准确率

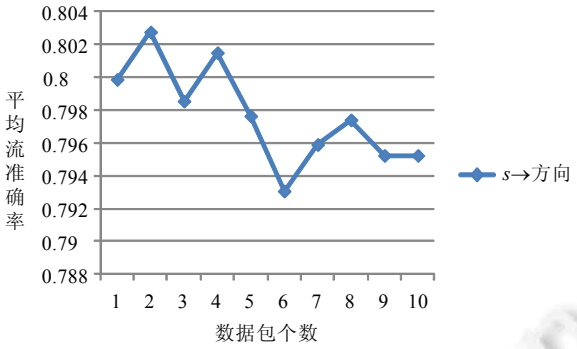


Fig.9 Overall accuracy achieved with the different number of packets in the server-to-client flows

图 9 服务器到客户端方向的流,使用不同的包数目获得的整体准确率

4.4 TCFEC与基于传统*K*均值聚类的流量分类器比较

本节从流准确率、字节准确率以及识别每个应用的检全率这 3 种度量来比较 TCFEC 与基于传统 *K* 均值聚类的流量分类器性能.

4.4.1 流准确率和字节准确率的比较

对于从客户端到服务器方向的流识别而言,由于仅观察流的前 3 个数据包,因此,TCFEC 由 3 个基分类器构成.其中,

- 第 1 个基分类器使用每个流的第 1 个数据包大小训练而成;
- 第 2 个基分类器使用每个流的前两个数据包大小训练而成;
- 第 3 个基分类器使用每个流的前 3 个数据包大小训练而成.

就客户端到服务器方向的流识别而言,由于 TCFEC 使用了前 3 个数据包大小,因此基于传统 *K* 均值聚类的流量分类器也使用前 3 个数据包大小(表 2 中表示为 *k*_means_3).同理,就服务器到客户端方向的流识别而言,TCFEC 使用了前两个数据包大小,因此基于传统 *K* 均值聚类的流量分类器使用前两个数据包大小进行流量分类(表 2 中表示为 *k*_means_2).

Table 2 Comparison of flow accuracies and byte accuracies

表 2 流准确率与字节准确率的比较

	数据流方向	平均流准确率	流准确率的标准差	平均字节准确率	字节准确率标准差
TCFEC	*c*→*s*	0.936 9	0.020 6	0.964 8	0.021 1
	s→*c*	0.860 9	0.042 6	0.892 2	0.010 6
*k*_means_2	*s*→*c*	0.805 6	0.026 9	0.854 1	0.011 0
*k*_means_3	*c*→*s*	0.922 1	0.024 9	0.922 6	0.070 1

表 2 比较了 TCFEC 与基于传统 *K* 均值聚类的流量分类器流准确率和字节准确率.容易看出,就流准确率和字节准确率而言,无论从客户端到服务器方向还是从服务器到客户端方向,TCFEC 明显好于基于传统的 *K* 均值聚类流量分类器.这是因为 TCFEC 的每个基分类器是用 *K* 均值聚类算法训练而成,TCFEC 在多个基分类器间做了最优的决策,以最小的错误率选择某个基分类器的结果作为最终结果.

4.4.2 识别每个应用的检全率比较

对于从客户端到服务器方向的 TCP 流,TCFEC 与每个基分类器识别不同应用的检全率如图 10 所示.可见,TCFEC 识别应用的检全率明显高于 *k*_means_3.这是由于 TCFEC 以最小的错误率在每个基分类器间做出最优的选择,因此,TCFEC 对每种应用识别的检全率获得较高的结果.对于分类服务器到客户端方向的 TCP 流,结果类似(如图 11 所示).

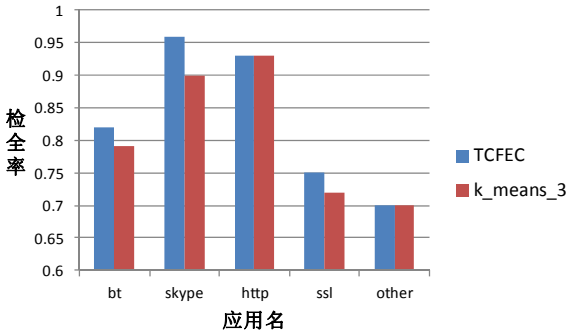


Fig.10 Comparison between TCFEC and *K*-means based classifier, when classifying client-to-server flows

图 10 分类客户端到服务器方向的流时, TCFEC 与 *K* 均值分类器的比较

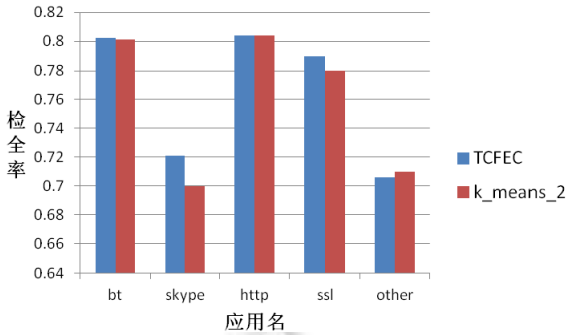


Fig.11 Comparison between TCFEC and *K*-means based classifier, when classifying server-to-client flows

图 11 分类服务器到客户端方向的流时, TCFEC 与 *K* 均值分类器的比较

4.5 TCFEC与典型流量分类算法的比较

我们前期的研究工作[23]使用了 Bagging 算法识别流量,Bagging 算法的基分类器为 C4.5 决策树.本节比较分析 TCFEC,SVM 和 Bagging 这 3 种算法的识别准确率.

4.5.1 流准确率和字节准确率的比较

表 3 对比了 TCFEC,SVM 和 Bagging 这 3 种分类算法的流准确率和字节准确率.可见,集成聚类 TCFEC 的识别结果明显要好于 SVM 和 Bagging 分类算法.更低的字节准确率标准差和流准确率标准差,表明在训练数据集发生变化时,集成聚类 TCFEC 分类更稳定.

Table 3 Comparison of flow accuracies and byte accuracies

表 3 流准确率与字节准确率的比较

	数据流方向	平均流准确率	流准确率的标准差	平均字节准确率	字节准确率标准差
TCFEC	$c\rightarrow s$	0.936 9	0.020 6	0.964 8	0.021 1
	$s\rightarrow c$	0.860 9	0.042 6	0.892 2	0.010 6
SVM	$c\rightarrow s$	0.890 3	0.062	0.954 6	0.054 1
	$s\rightarrow c$	0.805 5	0.060 7	0.835 2	0.046 2
Bagging	$c\rightarrow s$	0.902 5	0.051 1	0.945 3	0.039 6
	$s\rightarrow c$	0.851 6	0.043 9	0.867 7	0.021 0

4.5.2 比较每个应用的识别结果

图 12 和图 14 比较了 TCFEC,SVM,Bagging 分类器识别每一种应用的检全率.就客户端到服务器方向的 TCP 流识别而言(如图 12 所示),TCFEC 识别 bt,skype,ssl 的检全率明显高于 SVM 和 Bagging 分类算法;然而,TCFEC 识别 http 流量的检全率低于 SVM 分类算法.如图 12 所示,SVM 分类算法虽然识别 http 协议的检全率较高,但识别 http 协议的误报率却高达 50%.另一方面,就服务器到客户端方向的 TCP 流识别而言(图 14 所示),TCFEC 识别 bt,Skype 流量的检全率都高于 SVM 和 Bagging 分类算法.同样,如图 15 所示,Bagging 和 SVM 分类算法识别 http 流量的误报率也很高.由于 UNIBS 数据集中类别分布是不平衡的,http 流数占数据集中的大部分,而 SVM 和 Bagging 分类算法在不平衡的数据集上偏向于样本数较多的类别,因此,SVM 和 Bagging 分类算法识别 http 流数的误报率较高.但相对于 SVM 分类算法,Bagging 分类算法在不平衡训练数据集上的识别效果更好.

值得注意的是,由于从训练数据的预处理到流量分类器的设计,TCFEC 的实现充分考虑了流量分布的类别不平衡问题,对样本数较少的加密私有协议 Skype 流量识别的检全率较高且误报率较低.因此,TCFEC 更适合于识别加密的 Skype 流量.

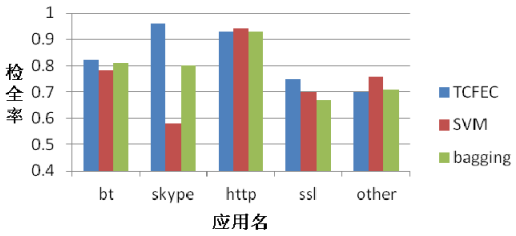


Fig.12 Comparison of TPRs when classifying client-to-server flows

图 12 分类客户端到服务器方向的流时,检全率比较

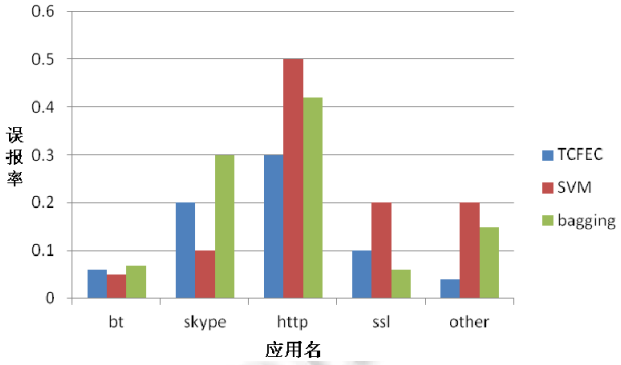


Fig.13 Comparison of FPR between TCFEC and SVM, when classifying client-to-server flows

图 13 分类客户端到服务器方向的流时,TCFEC 与 SVM 的误报率比较

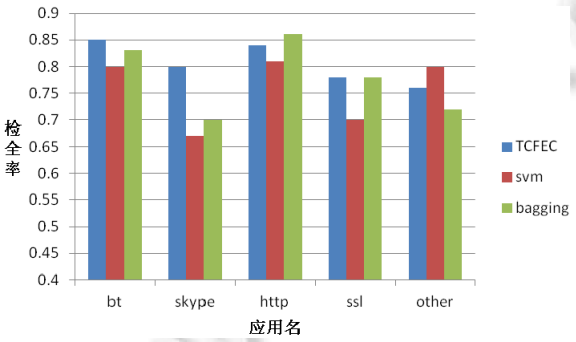


Fig.14 Comparison of TPRs when classifying server-to-client flows

图 14 分类服务器到客户端方向的流时,检全率比较

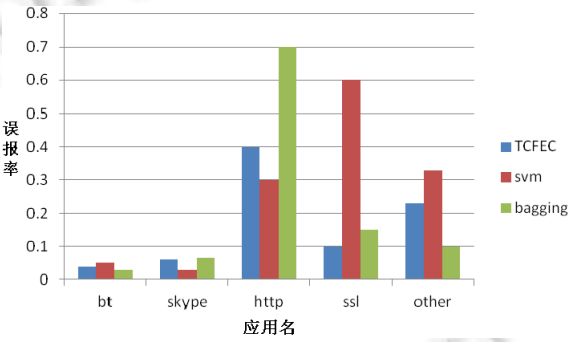


Fig.15 Comparison of FPR when classifying server-to-client flows

图 15 分类服务器到客户端方向的流时,误报率比较

5 总 结

本文提出了基于集成聚类的流量分类架构 TCFEC.TCFEC 仅需提取单向 TCP 流前若干个数据包大小作为特征,适合于在线分类流量.TCFEC 的每个基分类器通过在不同的特征子空间中聚类生成,对于分类不一致的样本,本文设计并实现了 SVM 决策器和 Hash 决策器以进一步决策该数据流的类别.通过与 K 均值聚类算法、SVM 分类算法和 Bagging 分类算法的比较,本文在公开的 UNIBS 数据集上验证了 TCFEC 的准确性和稳定性.

References:

[1] Zhang H, Lu G, Mahmoud TQ, Zhang Y, Yu XZ. Feature selection for optimizing traffic classification. Computer Communications, 2012,35(12):1457−1471. [doi: 10.1016/j.comcom.2012.04.012]

[2] Yang J, Wang Y, Qiao Y, Zhao X, Liu F, Cheng G. On evaluating multi-class network traffic classifiers based on AUC. Wireless Personal Communications, 2015,83(3):1731−1750. [doi: 10.1007/s11277-015-2473-4]

[3] Liu Q, Liu Z. A comparison of improving multi-class imbalance for internet traffic classification. Information Systems Frontiers, 2012,16(3):509−521. [doi: 10.1007/s10796-012-9368-7]

[4] Bernaille L, Teixeira R, Salamatian K. Early application identification. In: Proc. of the 2006 ACM CoNEXT Conf. New York: ACM Press, 2006. 1−12. [doi: 10.1145/1368436.1368445]

[5] McGregor A, Hall M, Lorier P, Brunskill J. Flow clustering using machine learning techniques. In: Proc. of the Passive and Active Network Measurement (PAM). LNCS 3015, Heidelberg: Springer-Verlag, 2004. 205−214. [doi: 10.1007/978-3-540-24668-8_21]

[6] Zander S, Nguyen T, Armitage G. Automated traffic classification and application identification using machine learning. In: Proc. of the IEEE Conf. on Local Computer Networks (LCN 2005). Sydney: IEEE Computer Society Press, 2005. 2257. [doi: 10.1109/LCN.2005.35]

[7] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms. In: Proc. of the 2006 SIGCOMM Workshop on Mining Network Data (MineNet 2006). New York: ACM Press, 2006. 281−286. [doi: 10.1145/1162678.1162679]

[8] Bernaille L, Teixeira R, Akodkenou I. Traffic classification on the fly. ACM SIGCOMM Computer Communication Review, 2006, 36(2):23−26. [doi: 10.1145/1129582.1129589]

[9] Erman J, Mahanti A, Arlitt M, Cohen I, Williamson C. Offline/Realtime traffic classification using semi-supervised learning. Performance Evaluation, 2007,64(9-12):1194−1213. [doi: 10.1016/j.peva.2007.06.014]

[10] Zhang J, Xiang Y, Zhou W, Wang Y. Unsupervised traffic classification using flow statistical properties and IP packet payload. Journal of Computer and System Sciences, 2013,79(5):573−585. [doi: 10.1016/j.jcss.2012.11.004]

[11] Zhang J, Chen C, Xiang Y, Zhou W, Vasilakos AV. An effective network traffic classification method with unknown flow detection. IEEE Trans. on Network and Service Management, 2013,10(2):133−147. [doi: 10.1109/TNSM.2013.022713.120250]

[12] Breiman L. Bagging predictors. Machine Learning, 1996,24(2):123−140. [doi: 10.1023/A:1018054314350]

[13] Li L, Hu Q, Wu X, Yu D. Exploration of classification in ensemble learning. Pattern Recognition, 2014,47:3120−3131. [doi: 10.1016/j.patcog.2014.03.021]

[14] Wang G, Sun J, Ma J, Xu K, Gu J. Sentiment classification: The contribution of ensemble learning. Decision Support Systems, 2014,57:77−93. [doi: 10.1016/j.dss.2013.08.002]

[15] Kuncheva L, Ludmila I, Dmitry P. Evaluation of stability of *k*-means cluster ensembles with respect to random initialization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(11):1798−1808. [doi: 10.1109/TPAMI.2006.226]

[16] Claesen M, Frank S, Suykens J. EnsembleSVM: A library for ensemble learning using support vector machines. Journal of Machine Learning Research, 2014,15:141−145.

[17] Szabo G, Szule J, Turanyi Z, Pongracz G. Multi-Level machine learning traffic classification system. In: Proc. of the 11th Int'l Conf. on Networks (ICN). Saint Gilles: IEEE, 2012. 69−76.

[18] Este A, Gringoli F, Salgarelli L. Support vector machines for TCP traffic classification. Computer Networks, 2009,53(14): 2476−2490. [doi: 10.1016/j.comnet.2009.05.003]

[19] Wright C, Monrose F, Masson G. HMM profiles for network traffic classification. In: Proc. of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security. Washington: ACM Press, 2004. 9−16. [doi: 10.1145/1029208.1029211]

[20] Dainotti A, Donato W, Pescape A, Salvo P. Classification of network traffic via packet-level Hidden Markov models. In: Proc. of the Global Telecommunications Conf. New Orleans: IEEE, 2008. 1−5. [doi: 10.1109/GLOCOM.2008.ECP.412]

[21] Este A, Gringoli F, Salgarelli L. On the stability of the information carried by traffic flow features at the packet level. ACM SIGCOMM Computer Communication Review, 2009,39(3):13−18. [doi: 10.1145/1568613.1568616]

[22] Nitesh V, Kevin W, Lawrence O, Philip K. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002,16:321−357.

[23] Zhang HL, Lu G. Machine learning algorithms for classifying the imbalanced protocol flows: evaluation and comparison. Ruan Jian Xue Bao/Journal of Software, 2012,23(6):1500−1516 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4074.htm [doi: 10.3724/SP.J.1001.2012.04074]

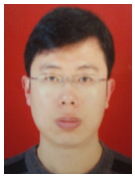
[24] Erman J, Mahanti A, Arlitt M. Byte me: A case for byte accuracy in traffic classification. In: Proc. of the 3rd Annual ACM Workshop on Mining Network Data (MineNet 2007). New York: ACM Press, 2007. 35−37. [doi: 10.1145/1269880.1269890]

[25] Yuan Z, Du C, Chen X, Wang D, Xue Y. SkypeTracer: Towards fine-grained identification for skype traffic via sequence signatures. In: Proc. of the Int'l Conf. on Computing, Networking, and Communications (ICNC). IEEE, 2014. 1−5. [doi: 10.1109/ICCNC.2014.6785294]

[26] Bonfiglio D, Mellia M, Meo M, Rossi D, Tofanelli P. Revealing Skype traffic: When randomness plays with you. ACM SIGCOMM Computer Communication Review, 2007,37(4):37–48. [doi: 10.1145/1282427.1282386]

附中文参考文献:

[23] 张宏莉,鲁刚.分类不平衡协议流的机器学习算法评估与比较.软件学报,2012,23(6):1500−1516. http://www.jos.org.cn/1000-9825/4074.htm [doi: 10.3724/SP.J.1001.2012.04074]



鲁刚(1982−),男,辽宁沈阳人,博士,工程师,主要研究领域为流量分类,网络行为分析与建模.

张宏莉(1973−),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为网络与信息安全,网络测量.

余翔湛(1973−),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为网络流量分类,网络行为分析.

郭荣华(1972−),男,博士,副研究员,CCF 专业会员,主要研究领域为网络流量分类,网络行为分析.