doi: 10.7690/bgzdh.2020.06.009

基于调整互信息的二进制协议一类分类算法

尹世庄[1]，王　韬[1]，陈庆超[1]，刘丽君[1]，张　斌[2]
（1. 陆军工程大学石家庄校区，石家庄 050003；2. 陆军工程大学南京校区，南京 210000）

摘要：为解决未知协议种类繁多、聚类结果不能涵盖所有协议的问题，提出一种基于调整互信息的一类分类算法。采用改进的 k-means 聚类，构建目标样本的合理覆盖模型，计算每个聚类中心的调整互信息阈值，得到各个聚类中心的调整互信息值，对不同聚类互信息阈值进行比较，并通过实验验证了算法的有效性。实验结果表明：与其他传统的一类分类方法相比，该方法在二进制协议一类分类中取得了较好的结果。

关键词：二进制协议；聚类中心；互信息；一类分类
中图分类号：TP391　　**文献标志码：**A

A Class of Classification Algorithm for Binary Protocol Based on Adjusting Mutual Information

Yin Shizhuang[1], Wang Tao[1], Chen Qinchao[1], Liu Lijun[1], Zhang Bin[2]
(1. *Shijiazhuang Campus of PLA University of Army Engineering*, *Shijiazhuang* 050003, *China*;
2. *Nanjing Campus of PLA University of Army Engineering*, *Nanjing* 210000, *China*)

Abstract: In order to solve the problem that there are many kinds of unknown protocols and the clustering results can not cover all protocols, a classification algorithm based on adjusting mutual information is proposed. By using the improved k-means clustering, the reasonable coverage model of the target sample is constructed, the adjusted mutual information threshold of each clustering center is calculated, the adjusted mutual information value of each clustering center is obtained, and the different clustering mutual information threshold is compared. The effectiveness of the algorithm is verified by experiments. The experimental results show that compared with other traditional classification methods, this method has achieved better results in the classification of binary protocols.

Keywords: binary protocol; clustering center; mutual information; a kind of classification

0　引言

协议分类是协议逆向的重要环节，协议种类繁多，分类器不可能包含所有协议的标签，因而在将未知协议输入分类器之前需要对协议进行初步筛选，判断其是否属于分类器的分类范围。一类分类[1]算法可以有效地解决这一问题。一类分类是指构建一个目标的覆盖模型，来判断测试样本是否属于模型的覆盖范围，主要包括密度估计法、聚类法、边界描述法 3 大类。其中聚类法对初始聚类中心及聚类个数的选取比较敏感，但是方法的复杂度低，如果能够准确地确定聚类中心，那么采用聚类法来进行一类分类还是比较有效的。一类分类器可以不用获取全部样本的标签，通过将样本的信息转换为高维空间的样本点，对分布模式进行识别；因此，笔者针对二进制协议一类分类问题展开研究，以促进模式识别在实际应用领域的发展。

1　相关理论和方法

1.1　基于聚类方法的一类分类

一类分类是指通过对训练目标样本进行学习，构建一个目标样本的合理覆盖模型，最后根据测试样本是否属于模型的覆盖范围实现分类鉴别。

密度估计方法的基本思想是通过参数化方法或非参数化方法，直接对训练目标样本集进行概率密度估计，并设置一个概率门限来决定目标类的决策边界。对于输入的未知样本，当输出概率高于所设置的门限时，则判定该未知样本为目标类，否则为非目标类。主要方法有高斯模型[2]、高斯混合模型[3]和 Parzen 窗函数法[4]。该类方法在高维样本有限的情况下，不能真实反映数据的本质特性，难以获得较好的分类性能。常见的聚类方法通过对一类样本的学习，也可以实现一类分类，比如 K 均值(Kmeans)和 K 中心(Kcenter)等[5]。该类方法通常假

收稿日期：2020-03-02；修回日期：2020-03-23
作者简介：尹世庄(1989—)，男，河北人，硕士，从事网络安全研究。E-mail: 18525742489@163.com。

定训练目标样本的分布满足某种聚类方式，通过聚类学习，以未知样本到最近簇类中心的距离判定该未知样本是否为目标样本。该类方法的运算复杂度小，但对聚类中心的初始化和聚类个数的选取非常敏感。

1.2 调整互信息

2 个随机变量的互信息[6-7](mutual information，MI)可以反映变量之间相互依赖性的程度。与相关系数不同的是，互信息适用范围更广，并不局限于实值随机变量。互信息也决定着分解的边缘分布的乘积 $p(X)p(Y)$和联合分布 $p(X,Y)$的相似程度[8]。其公式如下：

$$\mathrm{MI}(X;Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)。\tag{1}$$

式中，$p(x,y)$是 X 和 Y 的联合概率密度函数，$p(x)$和 $p(y)$分别为 X 和 Y 的边缘概率密度函数[9]。

从直观上看，互信息代表 X 和 Y 之间重叠的信息，也就是说，当知道 X 或者 Y 之中任意一个时，另外一个变量的不确定性减少的程度。MI 的取值范围是[0,1]，但是对于随机结果 MI 并不能保证值接近零，为解决这一问题，提出了调整互信息。

调整互信息[10](adjust mutual information，AMI)是互信息的一种改进，能够更好地反映数据分布的吻合程度。

$$\mathrm{AMI}=\frac{\mathrm{MI}-E\left|\mathrm{MI}\right|}{\max(H(U),H(V))-E\left|\mathrm{MI}\right|}。\tag{2}$$

式中：$H(U)$和 $H(V)$表示对应样本的边缘熵值；$E\left|\mathrm{MI}\right|$表示互信息的期望值。

利用基于互信息的方法来衡量聚类效果需要实际类别信息，AMI 的取值范围是[−1,1]，AMI 值越大，聚类结果与真实情况越吻合[11]。

2 基于调整互信息的一类分类模型

了检验二进制协议比特流是否属于已经聚好类的协议，采用调整互信息来衡量 2 条数据报文分布的相似程度。AMI 取值范围为[−1,1]，负数代表数据报文之间负相关，越接近于 1 的 2 个数据报文分布越相近。依据聚类的到的聚类中心首先计算聚类样本中每个类簇的样本到聚类中心的调整互信息阈值，然后计算测试数据中每条数据报文到各个聚类中心的调整互信息值，并判断是否满足任意一个类簇的阈值条件，若满足，则判定属于已经聚好类的协议；否，则属于完全未知协议，需要重新对这些数据进行聚类。

定义 1 类簇的平均调整互信息值 f。指的是某一聚类中心点到其他样本之间的调整互信息值之和除以样本数 n，$f_{\min}$ 是指判断是否属于该中心点类簇的最小调整互信息值。

$$f=\frac{\sum\mathrm{AMI}}{n};\tag{3}$$

$$f_{\min}=\frac{fk}{2}。\tag{4}$$

在某一类簇中，样本与中心点数据分布的吻合程度越高，调整互信息值越大。设定评价阈值 $f_{\min}$，其中 k 值为样本中聚类的个数。若样本的调整互信息值 I 大于任意一个评价阈值，则判定该样本属于已知聚类样本的协议种类。即若 $\exists I$

$$I_1>f_{1\min}\vee I_2>f_{2\min}\vee I_3>f_{3\min}\vee I_4>f_{4\min}\vee I_5>f_{5\min}。\tag{5}$$

则该协议报文属于已知聚类样本的协议种类。

提出了一种使用调整互信息的一类分类算法，算法的计算步骤如下：

参数：评价阈值 $f_{\min}$

输入：待分类的数据帧，k 个中心

输出：符合要求的样本

步骤：

输入训练样本

for 循环遍历每个中心点

for 循环遍历每条数据帧;

计算中心点与数据帧的调整互信息值(AMI)

将值赋予该数据帧的编号

计算 $f_{\min}$ 的值

输入测试样本

b=[]

n=[]

for i in range(len(I0)):

计算与各个聚类中心的调整互信息值;

if I0[i]＞f1 or I1[i]＞f2 or I2[i]＞f3 or I3[i]＞f4 or I4[i]＞ f5:

b.append(I[i])

n.append(i)

各类簇的调整互信息值，判定属于类簇的样本编号。

3 效果评估

在分类问题中，分类器有时能够正确地判断出

类别信息，有时却出现错误，所有可能的分类结果如表 1 所示。

表 1　分类器的分类结果

分类类别	目标类	非目标类
目标类	True Positive(TP)	False Negarive(FN)
非目标类	False Posirive(FP)	True Negarive(TN)

由表可知：实际正例数为 P=TP+FN，实际负例数为 N=FP+TN，实例总数为 Num=P+N。为了便于不同分类器间分类性能的比较，现概括总结几种常用的分类器性能评价指标：

1) 查准率，表示为测试目标样本正确识别数目与被识别为目标的测试样本数目的比值，即 precision=TP/(TP+FP)。

2) 查全率，表示为测试目标样本正确识别数目与测试目标样本总数的比值，即 recall= TP/(TP+FN)。

笔者使用正确识别率对结果进行评价。正确识别率是正确识别的目标类和非目标类之和除以总数。其公式为

$$\text{Accuracy}=\frac{(\text{TP}+\text{TN})}{\text{Num}}。\tag{6}$$

4　实验结果分析与评估

实验数据集由真实网络环境获取，协议是否已知并不影响聚类的准确度，为了更好地检验聚类效果，采用已知二进制协议代替未知协议。分别用 P_1、P_2、P_3、P_4 和 P_5 代表采集的 4 种未知二进制协议比特流子集：ARP、DNS、ICMP、TCP 和 SMB，样本量分别为 150 条。将这 750 条数据经过改进 k-means 算法聚类之后，打上标签作为训练数据集，并且提取出聚类结果的最终聚类中心。另重新抓取 300 条协议报文作为测试数据集，其中包括 ARP、DNS、ICMP、TCP、SMB、CDP 各 50 条。不失一般性，假定所有协议已近做了初步切分，数据报文均从对应协议头部开始并且包含数据部分。取最短数据报文长度 144 bit。每 4 bit 为一个处理单元，共 36 个特征。

为验证基于调整互信息的二进制协议一类分类算法的有效性，采用改进的 k-means 算法对第 4 组数据集进行聚类。通过实验得到 5 类聚类结果和 5 个最终的聚类中心。按照调整互信息聚类效果评估算法计算得到各类簇的调整互信息值。图 1 显示了测试样本与第 1 个聚类中心的调整互信息值的分布情况。

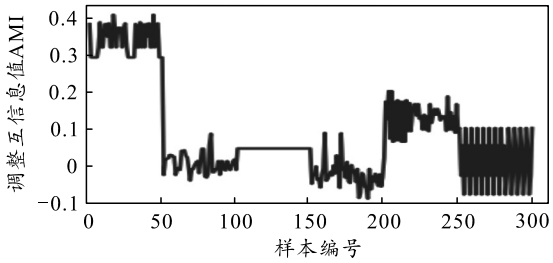

图 1　与第 1 个聚类中心的调整互信息值

图中样本编号为 X 轴，到对应聚类中心的调整互信息值为 Y 轴。各参数设置如下：第 1 个聚类中心值为[0, 0, 0, 1, 0, 8, 0, 0, 0, 6, 0, 4, 0, 0, 0, 1, 4, 3, 3, 6, 8, 8, 8, 11, 7, 5, 6, 7, 0, 4, 4, 6, 7, 2, 5, 4]，互信息阈值 $f_{1\min}$=0.269，调整互信息值反映了测试样本与第 1 个聚类中心的相似程度。经计算共 50 条协议符合要求。

同理计算第 2 个聚类中心与所有样本的调整互信息值，得到的结果分布如图 2 所示。各参数设置如下：第 1 个聚类中心值为[7, 6, 6, 7, 3, 0, 3, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 5, 6, 5, 6, 5, 6, 5, 5, 5, 6]，互信息阈值 $f_{2\min}$=0.260，经计算共 46 条协议符合要求。

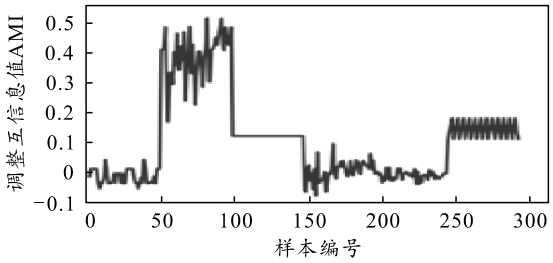

图 2　第 2 个中心互信息效果评估

同理计算第 3 个聚类中心与所有样本的调整互信息值，得到的结果分布如图 3 所示。各参数设置如下：第 1 个聚类中心值为[0, 5, 0, 0, 6, 6, 7, 7, 0, 4, 8, 8, 0, 0, 0, 2, 5, 14, 5, 12, 10, 5, 5, 9, 0, 0, 0, 0, 0, 0, 0, 0, 8, 6, 6, 7]，互信息阈值 $f_{3\min}$=0.384 8，经计算共 50 条协议符合要求。

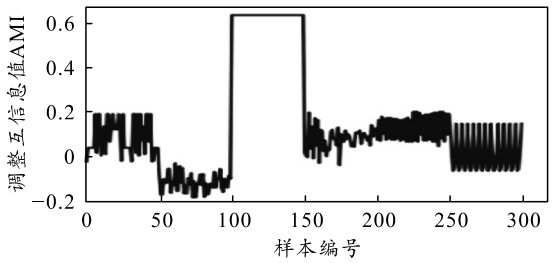

图 3　第 3 个中心互信息效果评估

同理计算第 4 个聚类中心与所有样本的调整互信息值，得到的结果分布如图 4 所示。各参数设置如下：第 1 个聚类中心值为[3, 1, 3, 2, 10, 4, 2, 7, 12, 4, 9, 6, 8, 6, 8, 7, 10, 3, 5, 6, 11, 3, 8, 7, 5, 0, 0, 0, 0, 1, 2, 2, 7, 6, 8, 8]，互信息阈值 $f_{4\min}$=0.053 6，经计算共 28 条协议符合要求。

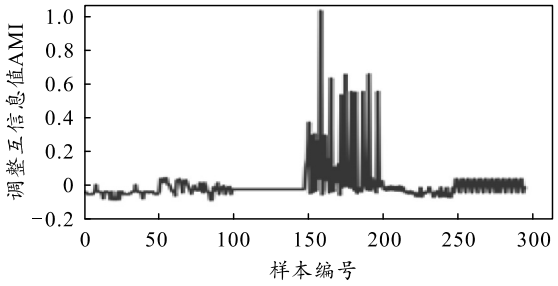

图 4 第 4 个中心互信息效果评估

同理计算第 5 个聚类中心与所有样本的调整互信息值，得到的结果分布如图 5 所示。各参数设置如下：第 1 个聚类中心值为[1, 5, 8, 7, 5, 10, 7, 7, 0, 0, 0, 0, 0, 0, 5, 5, 14, 14, 5, 3, 4, 12, 3, 2, 6, 2, 2, 6, 0, 0, 0, 0, 6, 0, 5, 8]，互信息阈值 $f_{5\min}$=0.270 6，经计算共 50 条协议符合要求。

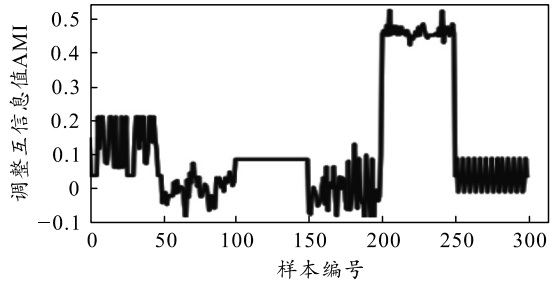

图 5 第 5 个中心互信息效果评估

图 6 反映了一类分类的结果。实验参数设置为：总共 300 条协议报文，其中 250 条为符合聚类标签要求的报文，50 条为未知报文。

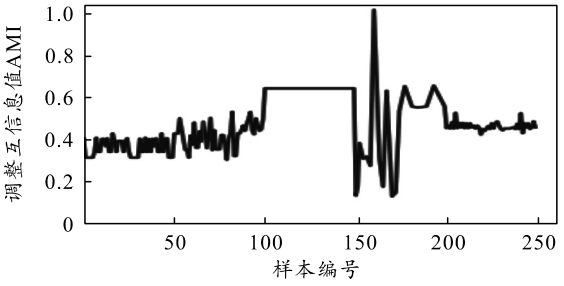

图 6 一类分类的结果

判定的结果为共 224 条满足结果，但是其中有 9 条重复，215 条正确，0 条错误。准确率 88.33%。相比于传统一类 SVM 分类器[12]，在同样数目的样本条件下，分类准确率高于传统一类 SVM 分类器的 86.4%。文中方法由于运算复杂度低，运算速度更快。

为了验证该方法在不同数据量下结果的有效性，分别取样本数据 300，480，600，900。分类的准确率如表 2 所示。

表 2 测试样本取不同值时分类算法的准确率

测试样本数据	300	480	600	900
准确率/%	88.33	82.56	87.64	89.76

表中准确率是采用不同测试样本分类后所得结果。从数据来看，测试样本的数目对分类结果的精度影响较小，数据本身的结构特点才是影响分类精度的重要因素。

通过对实验结果的分析可以看出：基于调整互信息的二进制协议一类分类算法是有效的，能对种类繁多的未知二进制协议进行一类分类，区分出经过聚类协议和完全未知的协议。

5 结束语

笔者从协议数据分布相似度出发，提出基于调整互信息的一类分类算法。该类方法的运算复杂度小。由于前期的 k-means 聚类是改进过聚类距离和初始聚类中心的改进算法，准确率达到了 95%以上，因而分类器的聚类中心是可靠的。通过实验验证了该方法的有效性，但在分类的过程中，如果 2 种协议的数据分布类似，而对应位置的语义不同，采用数理统计的方法很难区分，需要做进一步的研究。

参考文献：

[1] 李焕荣, 林健. 基于一类分类的聚类方法及其应用[J]. 计算机工程, 2005, 31(10): 36-38.

[2] 王洪德, 莫朝霞. 基于高斯模型的液氨储罐泄漏扩散仿真分析[J]. 中国安全科学学报, 2012, 22(9): 31.

[3] 白向峰, 李艾华, 李喜来, 等. 新型背景混合高斯模型[J]. 中国图象图形学报, 2011, 16(6): 983-988.

[4] LÁ Z M, HAYES M H, FIGUEIRAS-VIDAL A R. Training neural network classifiers through Bayes risk minimization applying unidimensional Parzen windows[J]. Pattern Recognition, 2018, 77: 204-215.

[5] 周本金, 陶以政, 纪斌, 等. 最小化误差平方和 k-means 初始聚类中心优化方法[J]. 计算机工程与应用, 2018, 54(15): 53-57.

[6] THOMAS J C R, PEÑAS M S, MORA M. New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance[C]//Chilean Computer Science Society, 2017.

[7] MAES F, COLLIGNON A, VANDERMEULEN D, et al. Multi-modality image registration by maximization of mutual information[C]//Workshop on Mathematical Methods in Biomedical Image Analysis. 1996.

[8] 茆诗松, 王静龙, 濮晓龙. 高等数理统计[M]. 北京: 高等教育出版社, 2006: 102−112.

[9] 谭丞, 李晓敏, 徐立军, 等. 基于联合概率密度判别器和神经网络技术的煤种辨识方法[J]. 机械工程学报, 2010, 46(18): 18−23.

[10] CHEN B, FAN Z Y, LI W M, et al. Holographic mutual information of two disjoint spheres[J]. Journal of High Energy Physics, 2018, 2018(4): 113.

[11] KRASKOV A, STÖGBAUER H, ANDRZEJAK R G, et al. Hierarchical clustering using mutual information[J]. Europhysics Letters (EPL), 2005, 70(2): 278−284.

[12] SCHÖLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the Support of a High-Dimensional Distribution[J]. Neural Computation, 2014, 13(7): 1443−1471.

****

（上接第 9 页）

[4] WANG L X, WU Z H, FU Y D, et al. Remaining life predictions of fan based on time series analysis and BP neural networks[C]//Information Technology Networking, Electronic and Automation Control Conference, IEEE. 2016: 607−611.

[5] MAIOR C B S, MOURA M D C, LINS I D, et al. emaining useful life estimation by empirical mode decomposition and support vector machine[J]. IEEE Latin America Transactions, 2016, 14(11): 4603−4610.

[6] 张国辉. 基于深度置信网络的时间序列预测方法及其应用研究[D]. 哈尔滨: 哈尔滨工业大学, 2017: 13−15.

[7] 吴昌友. 神经网络的研究及应用[D]. 哈尔滨: 东北农业大学, 2007: 6−11.

[8] JIANG X Y, LI S. BAS: Beetle antennae search algorithm for optimization problems[J]. International Journal of Robotics and Control, 2018, 1(1):1−5.

[9] JIANG X Y, LI S. Beetle antennae search without parameter tuning (BAS-WPT) for multi-objective optimization[J]. arXiv preprint arXiv, 1.

[10] 王甜甜, 刘强. 基于 BAS-BP 模型的风暴潮灾害损失预测[J]. 海洋环境科学, 2018, 37(3): 457−463.

[11] 查国清, 黄小凯, 康锐. 基于多应力加速试验方法的智能电表寿命评估[J]. 北京航空航天大学学报, 2015, 41(12): 2217−2224.

[12] 骆正山, 姚梦月, 骆济豪, 等. 基于 KPCA-BAS-GRNN 的埋地管道外腐蚀速率预测[J]. 表面技术, 2018, 47(11): 173−180.

[13] 张慰, 李晓阳, 姜同敏, 等. 基于 BP 神经网络的多应力加速寿命试验预测方法[J]. 航空学报, 2009, 30(9): 1691−1696.

[14] 裴洪, 胡昌华, 司小胜, 等. 基于机器学习的设备剩余寿命预测方法综述[J/OL]. 机械工程学报: 1-13[2019-05-21]. http: //kns. cnki. net/kcms/ detail/11. 2187. TH. 20190328. 2148. 120. html.

[15] 李静, 徐路路. 基于机器学习算法的研究热点趋势预测模型对比与分析: BP 神经网络、支持向量机与 LSTM 模型[J]. 现代情报, 2019, 39(4): 23−33.

****

（上接第 36 页）

3) 系统具有较强的抗干扰能力，适用于恶劣的工业环境，尤其是 HPM 系统脉冲功率源、强磁场工作时强辐射、强干扰的环境下，为 HPM 系统的测控系统的实现提供了一种可行的解决方案。在多种通信方式中，RS485 总线的抗干扰较强；ADAM 的各个功能模块在设计上采用输入输出隔离保护，抗干扰能力强；同时，现场和远程采用光纤的通信架构，远程均通过光信号传输，保证了系统运行的稳定性。

参考文献：

[1] 晋风华, 陈光大, 刘海英, 等. ADAM 模块在水电厂振动监测系统中的应用[J]. 国外电子元器件, 2002(4): 14−15.

[2] 周正贵. 基于 RS485 总线远程多点环境信息监测系统设计[J]. 长春师范大学学报, 2017, 36(6): 43−46.

[3] 孙庆玲. ADAM-4000 在 QCS 生料质量控制系统中的应用[J]. 建材技术与应用, 2006(5): 49−50.

[4] 潘昶. 基于多层协议字典的通用化通信技术[J]. 兵工自动化, 2019, 38(3): 31−34.

[5] 马云峰, 李培全, 唐述宏, 等. ADAM4000 系列模块在工业控制系统中的应用[J]. 微计算机信息, 1999, 15(3): 38−39.

[6] 薛加, 黄竞帆, 员来峰, 等. 弹药产品生产过程质量审核结果数据处理方法[J]. 兵器装备工程学报, 2019, 40(4): 128−131.

[7] 郭龙, 陈鸿, 李进杰, 等. 基于 ADAM 模块的航空训练模拟器数据采集与控制[J]. 现代电子技术, 2014, 37(18): 98−100.