·开发研究与工程应用· 文章编号:1000-3428(2018)05-0291-05 文献标志码:A 中图分类号:TP391

马尔科夫模型在网络流量分类中的应用与研究

赵 英[a],韩春昊[b]

(北京化工大学 a. 信息中心; b. 信息科学与技术学院,北京 100029)

摘 要: 传统的端口号与深度包检测分类技术已不能满足网络中各类应用的分类要求,无法进行准确分类。为此,提出一种基于半监督学习的马尔科夫模型网络流量分类算法。利用流之间的相关性构建马尔科夫模型,采用密度计算的方法估计聚类的中心点,通过 KL 距离计算中心点与样本之间的相似度,将样本划分到不同的应用类型中。使用马尔科夫模型提取特征参数,用以识别流量应用类型,并提高准确度,解决传统的基于半监督学习的流量分类方法依赖不稳定聚类算法的问题。实验结果表明,使用该方法机器学习得到的网络流量分类器可以取得理想的分类效果。

关键词: 网络流量分类;马尔科夫模型;相似度计算;半监督学习;流相关性;样本密度;聚类算法;相对熵

中文引用格式: 赵 英,韩春昊. 马尔科夫模型在网络流量分类中的应用研究[J]. 计算机工程,2018,44(5):291-295.

英文引用格式: ZHAO Ying, HAN Chunhao. Application and Research of Markov Model in Network Traffic Classification[J]. Computer Engineering, 2018, 44(5):291-295.

Application and Research of Markov Model in Network Traffic Classification

ZHAO Ying[a], HAN Chunhao[b]

(a. Information Center; b. College of Informatica Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

【Abstract】 With the development of information science and technology, the traditional port number and depth packet detection classification technology can not meet the classification requirements of various applications in the network, and can not be classified accurately. A Markov model network traffic classification algorithm based on semi-supervised learning is proposed. The Markov model is constructed by the correlation between flows. The center of the clustering is estimated by density calculation. The center point is calculated by KL distance. The similarity between samples is divided into different application types. The feature of the Markov model is used to identify the traffic application type and improve the accuracy. The problem of the traditional traffic classification method based on semi-supervised learning depends on the unstable clustering algorithm. Experimental results show that the network traffic classifier can achieve the ideal classification effect.

【Key words】 network traffic classification; Markov model; similarity calculation; semi-supervised learning; correlation of flow; sample density; clustering algorithm; relative entropy

DOI: 10.19678/j.issn.1000-3428.0046769

0 概述

网络流量分类是网络管理和网络安全的基础,是认识、管理、优化网络资源的重要依据。目前的流量分类技术主要基于端口号查询和深度包载荷检测的分类技术。但是,随着动态端口号和包载荷加密技术的应用的使用,2 种分类技术已经无法满足网络管理的需求。因此,近年来关于网络流量分类方法的研究主要是基于概率统计的机器学习算法研究。

机器学习在流量分类算法的运用中具有准确高、分类快速的优点,但其分类的好坏往往由训练集的选取决定。对于基于有监督学习的分类,识别能力取决于训练集样本中被标记的类型,如果测试集中出现训练集中不包括的类型,那么会影响识别精度;而基于无监督学习的分类器,它的分类精度完全依赖分类算法的好坏,但目前未出现能应对一切情况的算法。基于半监督学习的方法采用部分标记样本的方法构造分类器,结合了前 2 种方法的优点,提高了算法的准确度。但是,半监督算法中多数研究都采用如 k 均值这样的需要多次迭代的聚类算法,

基金项目: 中央高校基本科研业务费专项资金(PT1612)。

作者简介: 赵 英(1966—),男,教授、博士,主研方向为网络安全、并行计算;韩春昊(通信作者),硕士研究生。

收稿日期: 2017-04-13 **修回日期:** 2017-05-18 **E-mail:** hanch323@sina.cn

这类算法往往缺乏稳定的精度。

在网络流量分类研究中,文献[1]已经证实,使用对流量数据构造马尔科夫模型辅助分类,具有良好的准确性。但是怎样选取流量数据构建马尔科夫模型往往决定着流量分类结果的精度。而且以往的马尔科夫模型分类算法只能识别已知流量,当未知流量混入测试集样本中时,往往会严重影响分类的精度。

为此,本文提出一种基于网络流量相关性的马尔科夫模型,使用 KL 距离划分相似度较高的样本以形成类簇。由于以往的基于马尔科夫模型的分类器无法识别未知的流量类型,因此引入密度计算用以估计聚类中心点。

1 网络流量分类方法

1.1 流量数据采集

通常采集网络流量数据通过五元组作为最基本流量的特征,即源端口、目的端口、源 IP 地址、目的 IP 地址、协议,将具有相同五元组的流量数据称为流。具有五元组特征的流量数据通常传输在两台已接入网络的主机之间,而主机之间的数据传输是有方向性的。因此流还具有一下特性:

特性 1 流是具有传输在主机之间的单向有序流量数据的部分集合。

特性 2 流量采集时,凡传输时间超过 1 min 的流应当以 1 min 为单位被划分成不同的流。

根据马尔科夫模型构造的需要,通过五元组和流的特性采集流量数据样本。

1.2 马尔科夫模型

马尔科夫模型是指由多条马尔科夫链组成的模型,马尔科夫链的定义是由若干状态组成的随机序列,这样的序列中的状态量只与其前一个状态有关,称为无后效性,用公式表达如下:

$$P\{X_{n+1}=i_{n+1}\mid X_n=i_n,X_{n-1},\cdots,X_1=i_1\}=P\{X_{n+1}=i_{n+1}\mid X_n=i_n\} \tag{1}$$

式(1)表示了马尔科夫链的无后效性,决定第 n 项的只有第 $n-1$ 项状态,与 $n-1$ 之前的所有状态无关。马尔科夫模型是概率分布模型中所有可能的马尔科夫链的集合。

网络流量分类方法将具有相同应用类型的数据流分为一类。按照马尔科夫的定义和网络流的特征,并根据文献[2]中提出的前 4 个包已经足够以极高的准确率分类流量的观点,本文实验构造马尔科夫模型通过提取前 4 个包的大小[3]。定义马尔科夫模型中的状态量通过定义一个连续、有向的数据包,和包的大小。在 TCP 流量中,MSS(最大报文长度)是经常会发生变化的,如果直接以区间$[0,M_{\mathrm{MSS}}]$内的每一个整数作为一种状态,会造成状态过多而且很多状态并未出现,难以统计状态转移情况,因此,需要将状态重新归类。文献[4]提出将包的大小归类为 4 个区间,即[0,99],[100,299],[300,$M_{\mathrm{MSS}}-1$],[M_{MSS}],因为这些区间作为特征向量可以很好的区分多类应用。由于流的方向性,状态还可以被分类正向的和反向的,即客户端到服务器端流量归为正方向,服务器端到客户端为反方向,因此状态可以被分为 8 种,前 4 种代表客户端发往服务器的包,即{0,1,2,3},后 4 代表服务器发往客户端的包,即{4,5,6,7}。例如 0-1-2-3,是指客户端先发送[0,99]Byte 包,然后发送[100,299]Byte 包,接着发送[300,M_{MSS}]Byte 包,最后发送[M_{MSS}]包。

除了状态,通过统计和计算得出初始状态概率向量 $\boldsymbol{\pi}$ 和转换概率矩阵 $\boldsymbol{a}$:

$$\boldsymbol{\pi}_{\sigma_i}=\frac{F_0(\sigma_i)}{\sum_{j=1}^{n}F_0(\sigma_j)} \tag{2}$$

$$\boldsymbol{a}_{\sigma_i,\sigma_j}=\frac{F(\sigma_i,\sigma_j)}{\sum_{k}^{n}F(\sigma_i,\sigma_k)} \tag{3}$$

其中,F_0表示每种状态作为初始状态的个数,F 表示状态转换 i 到 j 的个数。

通过马尔科夫模型转化的网络流量概率分布模型,其优点在于通过计算每条马尔科夫链在分布模型中的概率,将一维的特征值(数据包大小)转化为多维特征参数,用各种数据流在应用类型中的分布情况来体现各类应用流量数据的特性,不需要选取过多的特征类型就可以体现出网络流量的分部特性。

1.3 流的相关性

在网络中的数据传输是以流的形式存在的,而流之间并不是独立存在的,是有相互关系的。根据文献[5]提出流之间的相关性可以表明网络流量的应用类型,具有相同的{dstIP,dstPort,protoType}属性的流属于同一类型,并经过试验取得较好的分类效果。因此,本文将具有流相关性的未知网络流量数据归位同一类型,为其构建马尔科夫模型。

2 基于马尔科夫模型的半监督聚类

2.1 算法问题描述

半监督学习流量分类方法结合了监督学习和无监督学习流量分类方法的特点,通常是先利用聚类算法在样本集中形成类簇,然后通过识别应用类别类簇中部分已标记样本来决定类簇的应用类型。半监督学习分类方法中的聚类方法通常选取如 K-means算法之类的需要多次迭代的方法,这些方法在样本集较为容易划分情况下迭代次数往往会小于分类样本个数,但是如果数据较难划分时迭代次数往往是不可控的。因此,使用迭代聚类算法用于网络流量分类往往稳定性较差。

在以往的网络流量分类研究中,马尔科夫模型

多用于监督学习分类方法[6]。文献[7]通过对流量数据构造可观测马尔科夫模型,通过似然分类器比较模型参数与已标记样本进行分类,得到了 95% 以上的准确率,因此,应用马尔科夫模型可以解决网络流量分类问题。但是以往的研究都存在相同的问题,就是当分类器用于实际环境流量分类时,由于网络技术的进步,新型流量层出不穷,因此无法对所有流量进行标记,基于马尔科夫模型的分类器会将未知流量误认为是已标记流量而造成错误分类。基于上述问题,主要解决的问题主要有以下两方面:

1)构造基于马尔科夫模型的半监督学习分类器,解决以往基于马尔科夫模型分类器无法识别未知流量的问题。

2)通过马尔科夫模型辅助聚类,以解决半监督分类中聚类算法稳定性问题。

2.2　相对熵

KL 距离(Kullback-Leibler Divergence)也叫相对熵(Relative Entropy),是评价相同事件空间中2 个概率分布的差异程度的量[8]。设样本 $X=\{x_1,x_2,\cdots,x_n\}$,$Y=\{y_1,y_2,\cdots,y_n\}$,那么 X 相对于 Y 的相对熵,即 KL 距离为:

$$H(X\parallel Y)=-\sum_{i=1}^{n}p(x_i)\lg p(x_i)-\left[-\sum_{i=1}^{n}p(x_i)\lg q(y_i)\right]=-\sum_{i=1}^{n}p(x_i)\lg\frac{p(x_i)}{q(y_i)} \tag{4}$$

其中,$p(x_i)$ 和 $q(y_i)$ 表示概率分布中样本 x_i 和 y_i 发生的概率[9]。使用马尔科夫模型进行聚类,需要将相似程度较高的样本聚集形成类簇,这就需要对样本相似度进行对比[10]。由于马尔科夫模型是概率分布模型,因此使用 KL 距离比较相似度。

2.3　聚类算法

对一个给定的样本集,如果其中某些样本点分布比较集中,那么这些点最有可能属于同一类型[11],那么这些分布集中的样本中,处于分布最密集区域的点最能代表该类型特性[12-13],因此可以通过求样本密度来选出这样的样本。设 $D_{\mathrm{KL}}(x_i,x_j)$ 表示样本 i 和样本 j 之间的 KL 距离,那么样本的密度可以定义为:

$$den\,s(x_i)=\frac{1}{m\sum_{j=1,j\neq1}^{n}D_{\mathrm{KL}}(x_i,x_j)} \tag{5}$$

密度计算式(3)表示样本距离周围的 m 个样本越近,则密度越大。大多数研究将 m 定义为所有样本的个数,这样会使样本密度受到较远距离样本的影响。因此采用邻域半径估计 m 的值,邻域半径 R 的定义如下:

$$R_i=\frac{\sum_{j=1,j\neq i}^{n}D_{\mathrm{KL}}(x_i,x_j)}{n^{\alpha}} \tag{6}$$

表示求总体样本集中样本 i 的平均距离,α 是调节参数。m 的值是所有 KL 距离小于 R_i 的样本点个数。算法执行前先对部分样本进行标记,算法描述如下:

1)设定聚类簇数为 k,已标记样本类型个数为 c。若 $k<c$,则说明样本中实际存在的类型个数多于类簇个数,这样会造成分类结果误差较大,算法结束,否则执行下一步。

2)初始化聚类中心点集合 $C=\{\cdot\}$。

3)设在已标记样本集中共标记 L 种应用类型,分别将每个已标记应用类型子集看做独立的的样本集,计算 L 个样本集中每个样本的邻域半径 R_i,根据 R_i 求得 m,然后求出样本密度 $dens(x_i,x_j)$,并从中选出密度最大的 L 个样本加入中心点集合 C 中,并从集合 S 中删除所有已标记样本。

4)若 C 中的样本个数小于 k,则从 S 中选出密度最大的一个样本,加入 C,并从 S 中将其删除,并删除其邻域半径内的样本。

5)迭代步骤 4),直至 C 中的样本个数等于 k。

6)输出集合 C。

经过上述步骤输出的 C 即为类簇中心点集合,利用 KL 距离计算其他样本和中心点的相似度形成类簇。

全部算法流程如下:

1)对所有样本集中每个样本按照流相关性划分成为若干个子集,每个子集中包含 N 个流,对每个子集构建马尔科夫模型,将形成马尔科夫模型的新样本放入集合 S 中。

2)使用 DPI 工具从样本集合 S 中取出部分样本进行标记。

3)通过密度计算获取中心点,从样本集合 S 中获取 k 个中心点样本构成集合 C,并从 S 中删除 C 中的样本。

4)对 S 中的样本根据中心点 C 使用 KL 距离进行聚类。

5)根据部分标记类型确定流量类型,根据分类结果构造分类器。

2.4　评价标准

为了正确评价各个相似度测量算法的分类结果,选用 Overall-accuracy 和 F-measure 作为评价指标[14-15]。用 TP、TN、FP、FN 分别表示真正例、真反例、假正例、假反例的样本个数,则上述评价标注的描述为:

$$O_{\mathrm{Overall\_accuracy}}(i)=\frac{\sum_{i=1}^{m}TP_i}{\sum_{i=1}^{m}TP_i+FN_i} \tag{7}$$

$$percision(i)=\frac{TP_i}{TP_i+FP_i} \tag{8}$$

$$recall(i)=\frac{TP_i}{TP_i+FP_i} \tag{9}$$

$$F_{\mathrm{F-meansure}}=\frac{2\times percistion\times recall}{percistion+recall} \tag{10}$$

3 实验结果与分析

3.1 实验环境及数据集

本文实验通过 TCP/IP 网络模型的分析对服务器端-客户端之间的通讯产生的流量进行采集,采集的数据集来自于以下的真实环境下的链路数据:BUCT 数据集。BUCT 数据集来自北京某大学网络实验室的节点路由上采集,可以获得全校人员访问网络的流量数据,共采集 182 GB 流量,约包含 4 245 173 个流。

为了准确的评价算法的准确性,本文实验使用基于深度包载荷检测工具(Ntop)和基于端口号采集工具(CoraReef),对数据集进行交叉验证判断其中的流量类型。对于 DPI 工具无法识别的加密流量,如 https 流量,使用端口号识别技术分类。使用手工检测分类工具无法识别的流量类型(约 100 000 个流),而这些流量大多数是新型 P2P 流量。最后去除 DPI 和手工检测均不能检测的流,利用其中的约 4 200 000 个流用于实验。经过检测这些流量中包含的的流量类型有 Web、SSH、SSL、FTP、Mail、P2P、Games,共 7 种流量类型。每类随机抽出 10 000 个流作为训练集样本,训练集包含共 70 000 个流。

3.2 实验结果

本文所提出算法中包含 2 个参数,即聚类簇数 k 和用来构建马尔科夫模型的流数 N,2 个参数的取值不同会对实验结果产生影响,因此,先测试在取不同 k 和 N 的情况下,分类结果的变化情况。

由图 1 可知,当 $N=10$ 时,所构建的马尔科夫模型基本上无法体现各个应用类型的特征,尽管 k 在增大,但是类型之间的区分度依然很差,所以准确度变化很小;当 $N=100$ 时,分类的准确率得到了较大提高,但是包含 100 个流的马尔科夫模型依然不能完整的表现类型的特性,因此,k 值的变化对准确性的影响较小;当 $N=300$ 时,可以从图中看出,马尔科夫模型已经能够表现完整的类型特征,k 值的变动对准确性也产生了较大影响。

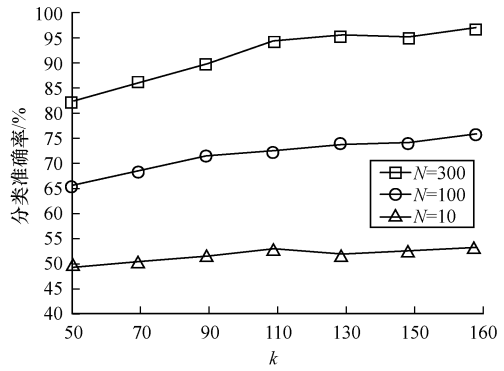

图 1 N 和 k 参数对分类准确度的影响

实验取 $N=\{10,100,300\}$,$k=\{50,70,90,110,130,150,160\}$。图 2 所示为分类的 Overall-accuracy 指标。

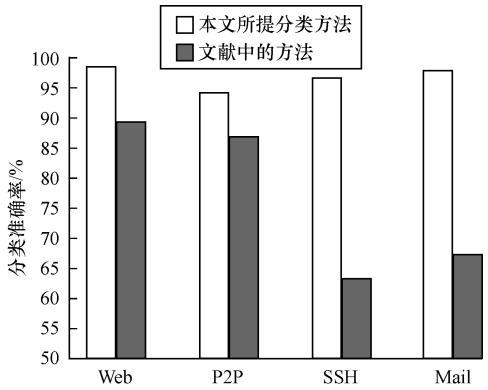

图 2 基于马尔科夫模型的分类器的准确性测试

然后进行第 2 组实验,即测试所提分类算法在测试集中存在未标记的类型存在时的准确性。选取部分 Web、Mail、SSH 和 P2P 流量进行标记,然后将这些样本和包含所有类型(包括 FTP、SSL 和 Game 流量)的部分未标记样本混合,训练半监督分类器。使用剩下的流量作为测试集,测试训练好的分类器和文献[2]提出的分类器的准确度。设定 $N=300$,$k=160$,分类结果如图 3 所示。

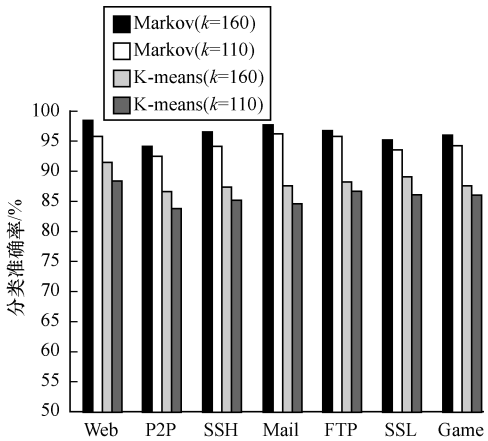

图 3 与 K-means 分类算法比较结果的 F-measure

对于文献所使用方法,由于未标记的类型的存在,且每种类型在样本集中的含量并不均匀,因此不同类型流量收到了不同程度的干扰。而本文所提方法影响较小,每一类型的准确率皆在 93% 以上,说明所提出的算法有识别未标记的流量应用类型的能力,实验结果达到了预期目标。

随后,测试所提出的算法与传统半监督算法在网络流量分类中的表现。选取基于半监督学习的 K-means聚类算法与其进行比较。选取能够表明流之间的相关性的三元组{dstIP,dstPort,protoType}作为 K-means 分类所需特征值。通过密度计算选择初始中心点,辅助 K-means 算法进行聚类。选取 $k=110$ 和 $k=160$作为 2 组被比较对象。

实验结果表明,本文所提出的分类方法与基于 K-means 的半监督分类算法相较,在准确度上有很明显的提升,使 Overall-accuracy 指标达到约 95% 。

能够有效提高分类的准确性主要由于以下因素:

1)利用马尔科夫模型推导出类型的参数进行分类识别,使应用类型的差异性得到了较好的反映。

2)通过流之间的相关性优化了马尔科夫模型的构建。

3)使用半监督学习分类方法使基于马尔科夫模型的分类器具有了识别未知流量类型的能力,在消除了未知流量干扰的情况下,基于马尔科夫模型的分类器的流量类型识别能力明显优于传统的聚类分类器。

4　结束语

本文研究马尔科夫模型在网络流量分类中的应用。使用密度计算估计聚类中心点,使马尔科夫模型分类器具有识别未知流量的能力,解决了传统基于半监督学习的分类器依赖不稳定聚类算法的问题。通过马尔科夫模型提取特征值,反映类型之间的差异性,提升半监督学习网络流量分类方法的稳定性,得到较高的精确度。实验结果证明了该算法的有效性。

参考文献

[1] MAIA J E B,HOLANDA F R. Internet traffic classification using a hidden markov model [C]//Proceedings of International Conference on Hybrid Intelligent Systems. Washington D.C.,USA:IEEE Press,2010:37-42.

[2] BERNAILLE L,TEIXEIRA R,SALAMATIAN K. Early application identification [C]//Proceedings of ACM Conference on Emerging Network Experiment and Technology. New York,USA:ACM Press,2006:1-12.

[3] FAHAD A,TARI Z,KHALIL I,et al. An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion [J]. Future Generation Computer Systems,2014,36(7):156-169.

[4] MÜNZ G,DAI H,BRAUN L,et al. TCP traffic classification using markov models[J]. Lecture Notes in Computer Science,2010,6003:127-140.

[5] 熊　刚,孟　姣,曹自刚,等. 网络流量分类研究进展与展望[J]. 集成技术,2012,1(1):32-42.

[6] ZHANG Jun,XIANG Yang,WANG Yu,et al. Network traffic classification using correlation information[J]. IEEE Transactions on Parallel and Distributed Systems,2013,24(1):104-117.

[7] FINAMORE A,MELLIA M,MEO M. Mining unclassified traffic using automatic clustering Techniques[C]//Proceedings of International Conference on Traffic Monitoring and Analysis. Berlin,Germany:Springer-Verlag,2011:150-163.

[8] 毕安琪,王士同. 基于 Kullback-Leiber 距离的迁移仿射聚类算法[J]. 电子与信息学报,2016,38(8):2076-2084.

[9] ZHANG Jun,XIANG Yang,WANG Yu,et al. A novel semi-supervised approach for network traffic clustering[C]//Proceedings of International Conference on Network and System Security. Washington D.C.,USA:IEEE Press,2011:169-175.

[10] FOREMSKI P. On different ways to classify internet traffic:a short review of selected publications[J]. Iitis PI,2013,25(2):119-136.

[11] PALMIERI F,FIORE U,CASTIGLIONE A. A distributed approach to network anomaly detection based on independent component analysis [J]. Concurrency & Computation Practice & Experience,2014,26(5):1113-1129.

[12] 周文刚,陈雷霆,Lubomir Bic,等. 基于半监督的网络流量分类识别算法[J]. 电子测量与仪器学报,2014,28(4):381-386.

[13] DAINOTTI A,DONATO W D,Pescape A,et al. Classification of network traffic via packet-level hidden markov models[C]//Proceedings of IEEE Global Telecommunications Conference. Washington D.C.,USA:IEEE Press,1930:1-5.

[14] 王　笑,李千目,戚　湧. 一种基于马尔科夫模型的网络安全风险实时分析方法[J]. 计算机科学,2016,43(s2):338-341.

[15] PARK J S,YOON S H,KIM M S. Performance improvement of payload signature-based traffic classification system using application traffic temporal locality[C]//Proceedings of Network Operations and Management Symposium. Washington D.C.,USA:IEEE Press,2013:1-6.

编辑　刘　冰

(上接第 285 页)

[12] BAEV I,RAJARAMAN R,SWAMY C. Approximation algorithms for data placement problems [J]. SIAM Journal on Computing,2008,38(4):1411-1429.

[13] MIRJALILI S,MIRJALILI S M,LEWIS A. Grey wolf optimizer[J]. Advances in Engineering Software,2014,69:46-61.

[14] 龙　文,赵东泉,徐松金. 求解约束优化问题的改进灰狼优化算法[J]. 计算机应用,2015,35(9):2590-2595.

[15] CHE X,IP B,LIN L. A survey of current YouTube video characteristics [J]. IEEE Multimedia,2015,22(2):56-63.

[16] FEI A,PEI G,LIU R,et al. Measurements on delay and hop-count of the Internet[EB/OL]. [2017-04-05]. http://nrlweb.cs.ucla.edu/nrlweb/publication/download/281/garyglcm98.pdf.

编辑　吴云芳