基于多标记与半监督学习的入侵检测方法研究

钱燕燕 李永忠 余西亚
(江苏科技大学计算机科学与技术学院 镇江 212003)

摘 要 机器学习所关注的问题是系统如何随着经验积累自动提高分类性能,这与入侵检测通过对外界入侵进行自我学习来提高其检测率和降低误报率是一致的。因此把机器学习的理论和方法引入到入侵检测中已成为一种有效方案。文中结合多标记与半监督学习理论,将ML-KNN算法应用于入侵检测系统。在KDD CUP99数据集上的仿真结果表明,该方法在入侵检测中能获得高检测率和低误报率。

关键词 多标记学习,ML-KNN算法,半监督学习,入侵检测
中图法分类号 TP393.08 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.029

Intrusion Detection Method Based on Multi-label and Semi-supervised Learning

QIAN Yan-yan LI Yong-zhong YU Xi-ya
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract The concerned problem of machine learning is how the systems automatically improve the classification performance with the increase of experience, which is consistent with IDS. Therefore, it has become an effective program to put the theories and methods of machine learning into IDS. In this paper, a multi-label lazy learning approach named ML-KNN was applied to intrusion detection systems. KDD CUP99 data set was implemented to evaluate the ML-KNN algorithm. The simulation results show that this method can achieve higher detection rate and lower false positive rate compared to other algorithms.

Keywords Multi-label learning, ML-KNN algorithm, Semi-supervised learning, Intrusion detection

入侵检测(intrusion detection)是一种动态的网络安全技术,能在系统受到危害之前拦截和响应入侵,提供了对内外部攻击的实时保护,是网络研究领域的热点[1]。入侵检测技术主要分为两类,即异常检测(abnormal detection)和误用检测(anomaly detection)。其中异常检测有检测新型攻击的能力且不需要入侵的先验知识,所以倍受入侵检测研究者的关注。异常检测的关键是如何建立系统的正常行为模式及如何利用该模式来检测和判断系统的异常行为。基于多标记学习的异常检测是一种无需指导的异常检测技术,它将相似的数据归为同一个聚类,将不相似的数据划分到不同的聚类。

在传统的监督学习框架中,待学习的样本具有明确的、单一的语义标记。在这个监督学习框架下已经提出多种入侵检测模型并取得了良好的效果。但是,现实世界的对象却通常具有多义性。例如,在文本分类中,一篇关于"杨威参加爸爸去哪儿"的新闻报道可能是娱乐报道,也可能是体育报道,也可能是地方报道,很难用单一的语义标记进行描述。于是,很自然的处理方式就是为一个文本赋予一个标记子集,并在此基础上建模和学习,这就构成了多标记的学习框架。在多标记学习框架下,样本由一个示例和对应的多个标记构成,学习的目标是将多个适当的标记赋予未知的示例。

多标记学习(multi-label learning)是机器学习中一个新的研究热点,基于多标记学习的算法已经成功应用于文档分类[2]、图像分类[3]、生物基因功能分类[4]等领域。本文用多标记学习算法对入侵检测数据集KDD CUP99进行分类并建立正常行为模式,再利用该模式来检测和判断系统的异常行为。实验证明,本文方法在一定程度上能够改善入侵检测的性能。

1 多标记学习

1.1 问题定义

假设 $X=R^d$ 代表 d 维示例空间,$Y=\{y_1,y_2,\cdots,y_q\}$ 表示 q 个类别的标记空间。传统的监督学习框架是单示例单标记学习[5-7],即一个对象只用一个示例来表示,而且该示例只对应一个类别标记。学习的主要目标是从数据集 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ 中学得函数 $f:X\to Y$,其中 $x_i\in X$ 为一个示例,而 $y_i\in Y$ 为示例 x_i 对应的类别标记。

而在多标记学习中,一个对象用一个示例来表示,而且该示例可以同时对应多个类别标记[4,7]。给定多标记训练集为 $T=\{(x_1,Y_1),\cdots,(x_m,Y_m)\}(x_i\in X,Y_i\in Y)$,其中 X 为输入空间,$Y=\{1,2,\cdots,Q\}$ 为有限个标记的集合。多标记学习系统的目标是从训练集 T 中进行学习,输出一个多标记分类器

到稿日期:2014-08-23 返修日期:2014-09-25 本文受江苏省高校自然科学基金项目(05KJD52006),江苏科技大学科研资助项目(2005DX006J)资助。

钱燕燕(1990—),女,硕士生,主要研究方向为网络与信息安全,E-mail:qyy19900227@163.com;李永忠(1961—),男,教授,主要研究方向为网络安全、计算机应用、藏文信息处理;余西亚(1983—),男,主要研究方向为极值统计。

$h: X \to 2^Y$。在一般情况下，为了得到上述的多标记分类器 $h(\cdot)$，学习系统将学习得到某个实值函数 $g: X \times Y \to R$。对于训练样本 x_i 及其对应的标记集 Y_i 而言，$g(\cdot,\cdot)$ 在属于 Y_i 的标记上输出较大的值，而在不属于 Y_i 的标记上输出较小的值，即 $y_1 \in Y_i$ 以及 $y_2 \notin Y_i$ 有 $g(x_i, y_1) > g(x_i, y_2)$。

1.2 问题转换

KDD CUP99 数据集[8]是较为权威的入侵检测数据集，除了包含大量的正常网络数据外，还包括属于 4 种攻击类型的 38 种不同的攻击：①Dos 拒绝服务；②Probe 扫描与探查；③R2L 未经授权的远程访问；④U2R 对本地超级用户的非法访问。KDD CUP99 数据集是个多标记学习问题。多标记学习目前的主要解决途径为：问题转化法和算法适应法。

(1)问题转换法的主要思想是通过对多标记训练样本进行处理，将多标记学习问题转换为其他已知的学习问题(如传统的单标记学习问题)进行求解，常用的两种方法是 BM(Binary Method)和 CM(Combined Method)。本文采用 BM 算法，假设多个标记之间是相互独立的，然后针对每一个标记，将数据集合按照包不包含当前，将问题转化为多个单标记问题。如对 KDD CUP99 数据集进行处理，得到单标记数据集合。数据集对应的样签如表 1 所列。

表 1 数据集对应的标签

标记			类型	标签
neptune	smurf	mailbomb	Dos	+1,−1,−1,−1
mscan	saint	satan	Probe	−1,+1,−1,−1
guess_password	snmpgetattack	snmpguess	R2L	−1,−1,+1,−1
buffer_overflow	httptunnel	ps	U2R	−1,−1,−1,+1

这里必须说明的是，每个攻击类型不止包括 3 个标记，本文选出的标记是每个类型中所占比例最多的 3 个。

(2)算法适应法的主要思想是通过改进传统的监督学习算法，使之能用于多标记数据的学习，代表性的算法有 ML_KNN[6]和 Rank-SVM[9]。

BM 算法虽然思想简单，易于实现，但是未考虑标记之间可能存在的其它关系。为了尽量弥补不足，本文结合以上两种方法，将算法 ML_KNN 应用于入侵检测，将各个分类看成多个二分类问题，以构造多标记学习系统，提高入侵检测性能。

2 半监督学习

在机器学习领域中，传统的学习方法有两种：监督学习和无监督学习。监督学习只能利用少量的有标签样本学习，而无监督学习只利用无标签样本学习[10]。半监督学习是监督学习与无监督学习相结合的一种学习方法，通过采用标签样本数据和未标签样本数据的联合概率分布来建立更好的学习方式。在标签的样本数据相对较少而未标签的样本数据相对较多的情况下，半监督学习在大多情况下都可以获得比监督学习和无监督学习更好的学习效果。

半监督学习分为半监督分类和半监督聚类，两者的差别在于，半监督分类是在监督分类的基础上，通过无标记数据指导分类过程，以提高分类的准备性；半监督聚类则是在无监督聚类的基础上，通过标记数据(或约束关系)指导聚类过程，以提高聚类质量[11]。本文主要围绕半监督聚类方法展开。

3 ML-KNN(Multi-Label K-Nearest Neighbor)算法

k 近邻(k-Nearest Neighbors，k-NN)算法的基本思想是：首先在训练样本集中找到与输入样本最近的 k 个邻居，然后用决策规则规定输入样本的类别。ML-KNN 算法[6,12]是对已有 k 近邻算法的改进。该算法的基本思想是采用“k 近邻”分类准则，统计近邻样本的类别标记信息，通过“最大化后验概率(Maximum a Posterior，MAP)”的方式推理无标签样本的所属集合[7]。

给定样本 x 和它对应的标记集合 $y \subseteq Y$，假定算法中一共取 k 个近邻。令 $\vec{y_x}$ 为对应样本 x 的标记向量，对于其中分量 $\vec{y_x}(s)(x \in Y)$，当 x 取得 s 时，$\vec{y_x}(s)=1$；否则 $\vec{y_x}(s)=0$。设 $N(x)$ 表示样本 x 在训练集中的 k 个近邻的集合，那么样本 x 的近邻中属于每个标记的数目组成的向量如式(1)所示：

$$\vec{C}_x(s)=\sum_{a\in N(x)}\vec{y}_a(s), s\in Y \tag{1}$$

对于每一个测试实例 t，ML-KNN 首先确定它在训练集合中的 k 个近邻组成的集合 $N(t)$。令 H_1^s 表示实例 t 中包含标记 s 的事件，H_0^s 表示实例 t 中不包含标记 s 的事件，并令 $E_j^s(j \in \{0,1,\cdots,k\})$ 表示在 t 的 k 个近邻中有 j 个包含标记 s 的事件。那么基于近邻标记计数向量 $\vec{C_t}$，测试样本 t 的预测标记向量 $\vec{y_t}$ 的计算公式如式(2)所示：

$$\vec{y_t}(s)=\arg\max_{b\in\{0,1\}} p(H_b^s \mid E^s_{\vec{C_t}(s)}), s\in Y \tag{2}$$

根据贝叶斯规则，式(2)可以重写为：

$$\vec{y_t}(s)=\arg\max_{b\in\{0,1\}}\frac{P(H_b^s)P(E^s_{\vec{C_t}(s)}\mid H_b^s)}{P(E^s_{\vec{C_t}(s)})}$$

$$\cong\arg\max_{b\in\{0,1\}} P(H_b^s)(E^s_{\vec{C_t}(s)}\mid H_b^s) \tag{3}$$

根据上面的公式可知，为了得到预测标记向量 $\vec{y_t}$，所需的信息包括先验概率 $P(H_b^s)(s\in Y, b\in\{0,1\})$和后验概率 $P(E_j^s \mid H_b^s)(j\in\{0,1,\cdots,k\})$，而这些信息都可以直接通过统计计算的方法从训练集合中得到。其算法伪代码描述如下：

```
%计算先验概率 P(H_b^s)
(1)for s∈Y do
(2)P(H_1^s)=(S+Σ_{i=1}^{m} y_{x_i}(s))/(S×2+m); P(H_0^s)=1−P(H_1^s);
   %计算后验概率 P(E_j^s | H_b^s)
(3)Identify N(x_i), i∈{1,2,…,m};
(4)  for s∈Y do
(5)    for j∈{0,1,…,k} do
(6)      c[j]=0; c'[j]=0;
(7)  for i∈{1,2,…,m} do
(8)    δ=C_{x_i}(s)=Σ_{a∈N(x_i)} y_a(s);
(9)  if (y_{x_i}(s)==1) then c[δ]=c[δ]+1;
(10)     else c'[δ]=c'[δ]+1;
(11) for j∈{0,1,…,k} do
(12) P(E_j^s | H_1^s)=(S+c[j])/(S×(k+1)+Σ_{p=0}^{k} c[p]);
(13) P(E_j^s | H_0^s)=(S+c'[j])/(S×(k+1)+Σ_{p=0}^{k} c'[p]);
     %计算 y_t 和 r_t
(14) 确定 N(t)
(15) for s∈Y do
(16) C_t(s)=Σ_{a∈N(t)} y_a(s);
(17) y_t(s)=arg max_{b∈{0,1}} P(H_b^s)P(E^s_{C_t(s)} | H_b^s);
```

(18) $\vec{r}_t(s)=P(H_1^s|E^s_{\vec{C}_t(s)})=(P(H_1^s)P(E^s_{\vec{C}_t(s)}|H_1^s))/P(E^s_{\vec{C}_t(s)})$

$=(P(H_1^s)P(E^s_{\vec{C}_t(s)}|H_1^s))/(\sum_{b\in\{0,1\}}P(H_b^s)P(E^s_{\vec{C}_t(s)}|H_b^s))$

其中，T 是训练数据集，k 是近邻的数目，t 是一个实例，$\vec{y}_t$ 就是输出的预测标记集合向量。此外，输入参数 s 是一个平滑参数，这里采用 Laplace 平滑参数，设 s 为 1。$\vec{r}_t$ 是一个实数向量，计算它的值用来对 Y 中的标记进行等级划分，其中 $\vec{r}_t(l)$ 对应的后验概率是 $P(H_1^l|E^l_{\vec{C}_t(l)})$。

4 实验与分析

为了研究多标记学习理论对入侵检测率性能的影响，实验采用了 KDD CUP99 中的 corrected. gz 数据集中的 9331 条数据记录，其分布如表 2 所列。

表 2 数据的类别分布

类别	Normal	Dos	Probe	R2L	U2R
数据	1815	6912	474	123	7
比例%	19.45	74.08	5.08	1.32	0.07

4.1 数据预处理

"0,udp,private,SF,105,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,253,0.99,0.01,0.00,0.00,0.00,0.00,0.00,0.00,snmpgetattack."为 KDD CUP99 数据集中的一条带有标签的数据。其中标签是数据的所属类别，在测试时仅作为判断条件。数据集中第 2、3、4 维属性都是符号型数据，为了使数据符合实验要求，需要对数据进行预处理。数据的预处理包括 2 个步骤：符号型数据的数据化处理和数值数据的标准化处理。对于符号型数据，对不同的符号取不同的数值，如第二维是网络协议类型，有 tcp、udp、icmp 3 种类型，分别取 tcp=1，udp=2，icmp=3，依此类推。为了消除不同的量纲对数据计算结果的影响，对数值型数据按式(4)进行标准差变换[13]，按式(5)进行极差变换归一化到[0,1]区间。

$$\overline{x_k}=\frac{1}{n}\sum_{i=1}^{n}x_{ik},\ s_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\overline{x_k})^2}$$

$$x'_{ik}=\frac{x_{ik}-\overline{x_k}}{s_k} \tag{4}$$

$$x''_{ik}=\frac{x'_{ik}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}}{\max\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}-\min\limits_{1\leqslant i\leqslant n}\{x'_{ik}\}},k=(1,2,\cdots,n) \tag{5}$$

4.2 实验结果与分析

利用半监督学习需要提供少量标记数据，因此人为地设计了 3 个数据集。其中，数据集一：标记数据约占总数据的 1/4；数据集二：标记数据约占总数据的 1/3；数据集三：标记数据约占总数据的 1/2。需要说明的是，3 个数据集含有相同的数据，但是标记数据所占的比例不同。本文方案是一种多分类算法，可以使用检测精度表示多分类算法的性能，同时可以查看半监督学习对性能的影响。检测精度的计算方法如下：

检测精度＝正确分类的样本数/总样本数×100%

本方案在数据集一、二、三上的检测精度结果如表 3 和图 1 所示，U2R 的训练和总样本数低，导致在数据集一中的检测精度为 0。给予更多训练样本后，在数据集二和三上的检测精度得到明显提高。而除了 U2R，其他攻击类型都有较好的检测精度，Dos 达到了 100%。由此可见，标记数据所占比例对最终的检测结果有一定的影响。

表 3 攻击类型的检测精度

测试数据	攻击类型 %			
	Dos	Probe	R2L	U2R
数据集一	100	97.74	97.83	0
数据集二	100	98.43	100	66.67
数据集三	100	99.58	100	100

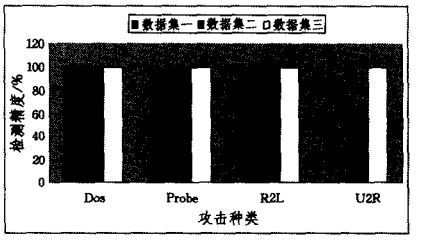

图 1 检测精度比较

事实上，入侵检测问题可以简化为二值假设检验问题，本文将 Normal 标记为(+1,−1)，将 Probe 、Dos、U2R 和 R2L 标记为(−1,+1)。因此可以用检测率(Detection Rate,DR)和误报率(False Positive Rate,FPR)进行描述：

检测率(DR)＝正确检测出的入侵数目/总的入侵数目×100%

误报率(FPR)＝将正常记录误判断为入侵记录数目/总的正常数目×100%

取不同 K 值，分别运行 10 次，实验结果如表 4 所列。以入侵检测的误报率和检测率作为衡量标准，尽量选择高检测率和低误报率下的实验数据。经权衡比较，本实验选择表 4、表 5、图 2、图 3 所示数据(加粗)作为实验结果。

表 4 不同 K 值下的检测性能

测试数据	ML-KNN 算法分类准确率(DR/FPR%)			
	K=1	K=3	K=5	K=7
数据集一	97.21/0.59	96.93/0.07	97.34/0.59	**98.35/7.48**
数据集二	**99.08/0.66**	96.87/0.16	98.36/3.63	97.41/2.64
数据集三	99.09/0.77	**99.09/0.33**	99.12/1.32	99.12/1.32

表 5 算法分类准确率比较(DR/FPR%)

	SK-Means	SFCA	本文
数据集一	81.34/66.88	97.98/44.54	98.35/7.48
数据集二	83.33/34.23	97.84/25.87	99.08/0.66
数据集三	87.52/45.71	98.98/29.85	99.09/0.33

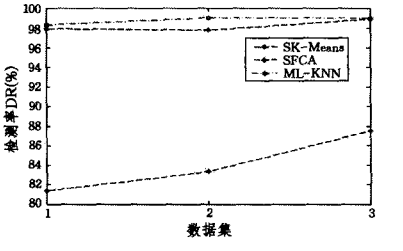

图 2 检测率比较

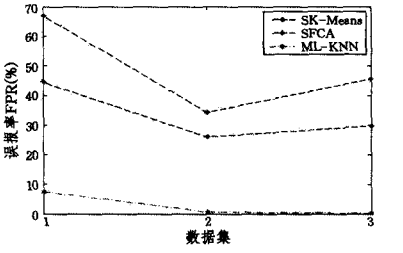

图 3 误报率比较

(下转第 146 页)

MCDH 假设下证明了方案的安全性。

参 考 文 献

[1] Sahai A, Waters B. Fuzzy identity-based encryption [M]// Advances in Cryptology-EUROCRYPT 2005. Springer Berlin Heidelberg, 2005: 457-473

[2] Yang P, Cao Z, Dong X. Fuzzy Identity Based Signature with applications to biometric authentication [J]. Computer Electrical Engineering, 2011, 37(4): 532-540

[3] Shaniqng G, Yingpei Z. Attribute-based signature scheme[C]// International Conference on Information Security and Assurance, 2008(ISA 2008). IEEE, 2008: 509-511

[4] Li J, Kim K. Hidden attribute-based signatures without anonymity revocation [J]. Information Sciences, 2010, 180(9): 1681-1689

[5] Li J, Au M H, Susilo W, et al. Attribute-based signature and its applications[C]// Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security. ACM, 2010: 60-69

[6] Escala A, Herranz J, Morillo P. Revocable attribute-based signatures with adaptive security in the standard model [M]// Progress in Cryptology-AFRICACRYPT 2011. Springer Berlin Heidelberg, 2011: 224-241

[7] Maji H K, Prabhakaran M, Rosulek M. Attribute-Based Signatures: Achieving Attribute-Privacy and Collusion-Resistance [OL]. http://eprint.iacr.org/2008/328.pdf

[8] Khader D. Attribute Based Group Signatures [OL]. http://eprint.iacr.org/2007/159.pdf

[9] Khader D. Attribute Based Group Signature with Revocation [OL]. http://eprint.iacr.org/2007/241.pdf

[10] Li J, Kim K. Attribute-Based Ring Signatures [OL]. http://epring.iacr.org/2008/394.pdf

[11] Li J, Huang Q, Chen X, et al. Multi-authority ciphertext-policy attribute-based encryption with accountability[C]// Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security. ACM, 2011: 386-390

[12] Shahandashti S F, Safavi-Naini R. Threshold attribute-based signatures and their application to anonymous credential systems [M]// Progress in Cryptology-AFRICACRYPT 2009. Springer Berlin Heidelberg, 2009: 198-216

[13] Maji H K, Prabhakaran M, Rosulek M. Attribute-based signatures[M]// Topics in Cryptology-CT-RSA 2011. Springer Berlin Heidelberg, 2011: 376-392

[14] Okamoto T, Takashima K. Efficient attribute-based signatures for non-monotone predicates in the standard model[M]// Public Key Cryptography-PKC 2011. Springer Berlin Heidelberg, 2011: 35-52

[15] Goyal V. Reducing trust in the PKG in identity based cryptosystems[M]// Advances in Cryptology-CRYPTO 2007. Springer Berlin Heidelberg, 2007: 430-447

(上接第 136 页)

一个好的入侵检测方案，不仅要尽量提高系统的检测率，而且要尽可能地降低系统的误报率，以提高系统报警的可信度。算法分类准确率比较如表 5 所列，相比于 SK-Means (semi-supervised K-Means)算法[14]和 SFCA(semi-supervised fuzzy clustering algorithm)算法[14]，ML-KNN 算法在标记数据的训练下建立了一个较好的模型。观察图 2 和图 3 可知，随着标记数据比例增加，算法的检测率逐渐提高，误报率明显降低；同时，ML_KNN 在检测率和误报率上明显优于算法 SK-Means 和 SFCA。因此将多标记和半监督学习应用于入侵检测，能够有效改善入侵系统的性能。

结束语 本文提出的方案具有更高的检测率和更低的误报率，实验证明，将多标记学习应用于入侵检测系统，能够很好地改善系统性能，优于传统的入侵检测算法。但本文算法是基于多标记学习 K-NN 算法，因此如何改进使其更适应入侵检测系统是目前有待解决的问题。同时本文对异常记录给予标记偏多，现实网络环境正常记录远远多于异常记录，如何模拟现实网络环境进行基于多标记学习的入侵检测实验也是今后值得考虑的研究方向。

参 考 文 献

[1] Wu Qing-tao, Shao Zhi-qing. Survey on intrusion detection techniques[J]. Application Research of Computers, 2005, 22(12): 11-44

[2] Schapire R E, Singer Y. Boostexter: A boosting-based system foe text categorization [J]. Machine Learning, 2000, 39(2/3): 135-168

[3] 宋相法，焦李成. 基于稀疏编码和集成学习的多示例多标记图像分类方法[J]. 电子与信息学报，2013，35(3)：622-626

[4] 陈晓峰，王士同，曹苏群. 半监督多标记学习的基因功能分析[J]. 智能系统报，2008，3(1)：83-90

[5] 周志华，张敏灵. MIML：多示例多标记学习[J]. 机器学习及其应用，2009：218-234

[6] Zhang Min-ling, Zhou Zhi-hua. A Lazy Learning Approach to Multi-Label Learning [J]. Pattern Recognition, 2007, 40(7): 2038-2048

[7] 周志华，杨强. 机器学习及其应用[M]. 北京：清华大学出版社，2011：179-199

[8] University of California, Irvine. KDD cup 1999 data[EB/OL]. 1999-10-28 [2012-03-20]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[9] Elisseeff A, Weston J. A kernel method for multi-labelled classification[C]// Dietterich T G, Becker S, Ghahramani Z., eds. Advances in Neural Information Processing Systems 14 (NIPS' 01). Cambridge, MA: MIT Press, 2002: 681-687

[10] 袁利永，王基一. 一种改进的半监督 K-Means 聚类算法[J]. 计算机工程与科学，2011，33(6)：138-143

[11] 夏战国，万玲，蔡世玉，等. 一种面向入侵检测的半监督聚类算法[J]. 山东大学学报：工学版，2012，42(6)：1-7

[12] 郭跃健，李宏. 多值属性和多标记数据分类[D]. 长沙：中南大学，2010

[13] 谢中华. Matlab 统计分析与应用：40 个案例分析[M]. 北京：北京航空航天大学出版社，2010

[14] 王汝山，李永忠. 基于半监督聚类的入侵检测技术研究[D]. 镇江：江苏科技大学，2010