基于距离加权模板约简和属性信息熵的增量SVM入侵检测算法

徐永华 李广水

(金陵科技学院信息技术学院 南京 211169) (江苏省信息分析工程实验室 南京 211169)

摘 要 为了解决SVM入侵检测方法检测率低、误报率高和检测速度慢等问题,提出了一种基于距离加权模板约简和属性信息熵的增量SVM入侵检测算法。该算法对K近邻样本与待测样本赋予总距离加权权重,对训练样本集进行约简,并以邻界区分割和基于样本属性信息熵对聚类样本中的噪声点和过拟合点进行剔除,以样本分散度来提取可能支持向量机,并基于KKT条件进行增量学习,从而构造最优SVM分类器。实验仿真证明,该算法具有较好的检测率和检测效率,并且误报率低。

关键词 入侵检测,SVM,距离加权,信息熵,邻界区

中图法分类号 TP393.08 **文献标识码** A

Incremental SVM Intrusion Detection Algorithm Based on Distance Weighted Template Reduction and Attribute Information Entropyc

XU Yong-hua LI Guang-shui

(School of Information Technology, Jinling Institute of Technology, Nanjing 211169, China)

(Jiangsu Information Analysis Engineering Laboratory, Nanjing 211169, China)

Abstract In order to solve the problem of the SVM intrusion detection method which has low detection rate, high distorting rate and slow detection speed, a kind of incremental SVM intrusion detection algorithm based on distance weighted template reduction and the attribute information entropy was proposed. In this algorithm, the training sample set reduction is made according to the sample for the samples and the neighbors to the total distance weighted weight, then, the clustering sample point and the noise of the fitting point are taken out through the adjacent to the border area segmentation and based on sample attribute information entropy, and then, using the sample dispersion extracts possible support vector machine, and incremental learning based on KKT conditions is made to construct the optimal SVM classifier. The simulation results show that the algorithm has good detection rate and the detection efficiency, and distorting rate low.

Keywords Intrusion detection, SVM, Weighted distance, Information entropy, Adjacent to the border area

1 引言

入侵检测是网络安全防御的关键技术,将具有较强的自学习性、适应性和鲁棒性的机器学习算法应用于入侵检测系统,可在提高检测系统检测率的同时降低误报率。因此基于机器学习的入侵检测方法研究极其重要[1]。支持向量机SVM(Support Vector Machine)作为一种基于核函数的学习方法,可利用核函数将不可分的非线性训练集映射到高维特征空间,使其在线性可分,并尽可能地构造具有最大间隔的超平面作为最优分类面,从而对样本数据集进行分类和识别[2]。支持向量机的分类函数是高位空间的超平面,SVM学习的过程就是对超平面进行优化调整的过程。文献[3]提出了一种基于距离加权的模版约简K近邻算法,其在对训练数据集进行分类时将边界样本去掉,以近邻与待测样本的距离为度量赋予不同的权重,但是的近邻样本权重易陷入局部最大权重;文献[4]提出了基于属性信息熵的K邻近算法,该算法对样本信息熵与待测样本距离最小的k个近邻进行分类度量,但是对于样本集属性平均信息熵的计算方法有待改进;文献[5]提出了一种增量式SVM入侵检测方法,其通过K均值聚类方法对训练集进行训练,以邻界区分割和最优超平面来实现SVM分类器的构造,但是该方法中的聚类计算对初始值较为敏感,且对训练样本集没有进行约简,计算量较大。基于以上研究,提出了基于距离加权模板约简和属性信息熵的增量SVM入侵检测算法,其通过约简和多次优化构造最优超平面,并进行增量学习,从而构造最优SVM分类器。实验证明,该算法在保证检测率的前提下降低了误报率,提升了检测效率。

到稿日期:2012-02-22 返修日期:2012-07-23 本文受江苏省教育厅高校自然科学研究项目(11KJD510002)资助。

徐永华(1971—),男,硕士,讲师,主要研究方向为数据挖掘与计算机网络,E-mail:xyh@jit.edu.cn;李广水(1965—),男,博士,教授,主要研究方向为数据挖掘、服务计算。

2 基于距离加权模板约简和属性信息熵的SVM入侵检测算法

2.1 基于距离加权的模板约简K近邻聚类分析

定义1(最近邻链[6]) 类别 w_i 的样本 x_i 的最近邻链可表示为：

$$\begin{cases} x_{i0} & x_{i1} & \cdots & x_{ik} \\ d_{i0} & d_{i1} & \cdots & d_{ik} \end{cases}$$

其中，样本序列 x_{ik} 的起始样本为 $x_{i0}=x_i$，当 $x_{i,k+1}=x_{i,k-1}$ 时，序列终止于 x_{ik}，且 x_{ij} 为 $x_{i,j-1}$ 的最近邻，若 j 为偶数，则 x_{ij} 属于 ω_i，否则不属于该类别。其中距离序列的距离 d_{ij} 是样本 $(x_{i,j+1},x_{ij})$ 间的欧氏距离，$d_{ij}=\sqrt{(x_{i,j+1},x_{ij})^2}$，且 $d_{ij}\geqslant d_{i,j+1}$。

设 $L=\{(w_m,x_i)\mid i=1,2,\cdots,n,m=1,2,\cdots,c\}$ 为由 n 个样本构造成的训练集，其 c 个类别对应的样本类别 ω_m 已知，样本 x_t 和其类别 ω_t 待测，则基于距离加权的模板约简算法为：

Step1 设 α 为约简阈值，当有限样本集存在临界值 N，且 $\alpha\geqslant N$ 时，样本集中不存在约简样本。α 的取值会影响样本约简的精度，一般是取初始值，可通过多次约简，调整取值。

Step2 对训练集 L 中的样本 x_i 求其对应的最近邻链 C_i。

Step3 对于 C_i 的距离序列，若

$$d_{ij}>\alpha * d_{i,j+1},j=\begin{cases}0,2,\cdots,l-3, & l\text{为奇数}\\ 0,2,\cdots,l-2, & l\text{为偶数}\end{cases}$$

则标记该样本 x_{ij}。

Step4 对训练集 L 中标记的样本 x_{ij} 进行约简删除，得到样本 L'。

Step5 当样本为多类分类问题时，对训练集 L 的任意两类样本都进行两类问题的算法约简，则对于 c 类问题，每个样本进行 $c-1$ 次约简，得到 $c-1$ 个约简集合，并将同一类别的约简结果进行合并，作为该类别的最终约简结果。

Step6 对于待测样本 x_t，在训练集 L' 中查找与之最近邻的 k 个样本，且 $\{x_{t1},x_{t2},\cdots,x_{tk}\}$ 对应的类别为 $\{\omega_{t1},\omega_{t2},\cdots,\omega_{tk}\}$，求最近邻时用欧氏距离作为度量，则待测样本 x_t 的距离序列为 $\langle d_{t1},d_{t2},\cdots,d_{tk}\rangle$。

Step7 利用反距离加权为 k 个近邻样本的贡献度赋予相应的样本权值，则第 j 个近邻样本的权值可表示为：$w_{tj}=\frac{1}{d_{tj}^2},j=1,2,\cdots,k$，而对于待测样本 x_t 的距离序列 $\langle d_{t1},d_{t2},\cdots,d_{tk}\rangle$，其与训练集 L' 中的样本 x_i 的距离为 $d_{(x_t,x_i)}$，则 $w_{ti}=\frac{d(x_t,x_i)^{-\beta}}{\sum_{i=1}^{k}d_{ti}^{-\beta}}$。$\beta$ 为指数值，其取值与样本点和待测点间的距离相关。

Step8 则对于待测样本集 x_t，其判别函数 f 为：

$$f(\omega_m,x_t)=\sum_{i=1}^{k}w_{ti} * I(\omega_m=\omega_{ti})\quad m=1,2,\cdots,c$$

式中，

$$I=\begin{cases}1, & \omega_{ti}=\omega_m\\ 0, & \omega_{ti}\neq\omega_m\end{cases}$$

则将最大的 $f(\omega_m,x_t)$ 的类别 ω_m 标记为待测样本 x_t 的类别，且有

$$\omega_t=\operatorname{argmax}_{\omega_m}f(\omega_m,x_t)$$

2.2 邻界区的分割

将正负两类样本集分别进行基于距离加权的模板约简 K 近邻聚类分析，设 $\{C_+,C_-\}$ 为正负两类样本的子聚类集合，$\{V_+,V_-\}$ 表示其子聚类中心集合。

定义2(子聚类中心距离矩阵) 设第 i 个正类中心与第 j 个负类中心样本点的欧氏距离为 D_{ij}，则 $D=(D_{ij})_{C_+\wedge C_-}$ 为子聚类中心的距离矩阵。

定义3(邻界子聚类区) 设子聚类中心距离矩阵 D 的正类与负类中心存在距离最近的 k 个子聚类中心，则称这 k 个子聚类中心的子聚类区为正类样本的 k-邻界子聚类区 $Z(v_+^i)$。

设正类的邻界区为 B_+，则 $B_+=Z(v_-^1)\vee Z(v_-^2)\vee\cdots\vee Z(v_-^c)$。

负类的邻界区 B_- 为：$B_-=Z(v_+^1)\vee Z(v_+^2)\vee\cdots\vee Z(v_+^{c_+})$。

2.3 基于属性信息熵的分类超平面构造

将邻界区中的所有样本与分类超平面 $g=0$ 的距离 d_t 进行对比，距离较远的样本点成为支持向量机向量的概率较小，因此参考文献[6]的方法对其筛选，则：

以正类样本集为例，设 S_i 为 B_+ 的子聚类样本集 i 包含的样本数目，属性 V 有 i 个不同的取值，其属性 v_i 在 B_+ 中出现的次数为 $|v_i|$，且属于第 j 类样本集的个数为 $|v_{ij}|$，则属性的信息熵为：

$$H(v_i)=-\sum_{j=1}^{n}p_{ij}\ln p_{ij}$$

式中，$p_{ij}=\frac{|v_{ij}|}{|v_i|}$，是属性为 v_i 的样本为类别 ω_i 的概率，且当 $|v_i|=|v_{ij}|$ 时，$H(v_i)=0$。

则对正类样本集中的任意两个样本 (x_i,x_j)，其属性值均为 $\{v_1,v_2,\cdots,v_n\}$，则 x_i 与 x_j 的距离为：

$$d(x_i,x_j)=\frac{1}{n}\sum_{i=1}^{n}H(v_i)$$

若 (x_i,x_j) 为正类样本集的子聚类中心，则 $d(x_i,x_j)$ 为子聚类中心的信息熵距离，对于具有 c 个子聚类中心的正类样本集 B_+，其间类平均信息熵距离 $\overline{d_{ic}}$ 为：

$$\overline{d_{ic}}=\frac{1}{c-1}\sum_{k=1}^{c-1}d(x_{ik},x_{jk})$$

则对正类样本集中距最优分类面 $g^*=0$ 距离较远的 d_{ti} 与 $\overline{d_{ic}}$ 进行对比，若 d_{ti} 大于 $\overline{d_{ic}}$，则抛弃该样本点；否则保留该样本点。

2.4 支持向量机的提取

通过属性信息熵的分类超平面构造是最优超平面的近似，可能会存在噪声数据或过拟合样本，因此需要对支持向量机进行进一步提取。

定义4(正类样本分散度) 设正类样本到分类超平面 $g=0$ 的欧氏距离均值为 d^+，样本 (x_i,x_j) 与 $g=0$ 的距离和 d^+ 的差值为 $d_{v_i}^+$，则

$$Degree^+=\frac{1}{\sum_{i=1}^{c_i^+}(d_{v_i}^+)^2}$$

为正类样本分散度，其中 c_i^+ 为正类样本总数目。

则负类样本分散为：

$$Degree^{-}=\frac{1}{\sum_{i=1}^{C_i^{-}}(dv_i^{-})^2}$$

定义5(支持向量) 设 ω 为类分类阈值,则正类样本 (x_i,x_j) 到中心分类超平面的距离 $d(x_{ij})$ 为:

$$|d(x_{ij})-\frac{\omega}{2}|\leqslant Dregree^{+}*\sigma^{+}$$

式中,α^{+} 的值域为(0,0,5)。

则负类样本的可能支持向量判别条件为:$|d(x_{ij})-\frac{\omega}{2}|\leqslant Dergree^{-}*\sigma^{-}$,$\alpha^{-}\in(0,0.5)$。

将定义5中的可能正负两类支持向量机合并,形成支持向量机 SV。通过该条件可将噪声样本和过拟合样本点剔除,有利于提高 SVM 的分类精度。

2.5 基于 KKT 条件的增量学习

增量样本集设为 LDB,利用经过支持向量机构造的分类超平面 $g'=0$ 和 KKT 条件对其进行分类,则违反 KKT 条件的增量样本主要有[7]:

1)若 $0\leqslant d(x_i)\leqslant 1$,则样本被分类在分类间隔中的本类点集;

2)若 $-1\leqslant d(x_i)<0$,则样本被分类在分类间隔中的异类点集;

3)若 $d(x_i)<-1$,则样本被分类在分类间隔外的异类点集。

设 TB 为 LDB 中违反 KKT 条件的样本点集,对 LTB 进行增量学习,可得到新的支持向量机 SV^2 和分类超平面 $g^2=0$。则其迭代增量学习算法如下:

Step1 通过 SV^1 构造分类超平面 $g'=0$;

Step2 若 $LTB\neq\emptyset$,对 LTB 进行增量学习,确定违反 KKT 条件的样本集 T;

Step3 若 $T=\emptyset$,则 $SV^2=SV^1$,$g^2=g^1$,转 Step2;否则转 Step4;

Step4 $SV=SV\cup T$,转 step2。

2.6 算法描述

基于距离加权模板约简和属性信息熵的 SVM 入侵检测算法为:

输入:正类和负类训练样本$\{DB^{+},DB^{-}\}$,聚类个数 k,属性 v_i,参数$\langle\alpha,\beta,\sigma^{+},\sigma^{-}\rangle$

输出:最优超平面 $g^{*}=0$

Step1 利用基于距离加权的模板约简 K 近邻聚类分析算法,对正负两类训练样本集进行聚类分析;

Step2 根据定义2计算正负类样本的聚类中心距离,构造类间子聚类中心距离矩阵 D;

Step3 根据定义3计算正负两类子聚类中心的 k-邻界子聚类区,并构造邻界区 B;

Step4 利用基于属性信息熵的分类超平面构造算法对邻界区 B 中的样本集进行筛选,得到分类超平面 $g=0$;

Step5 利用定义4对正负样本的分散度进行计算,依据定义5进行支持向量机的提取;

Step6 利用基于 KKT 条件的增量学习得到最优分类超平面 $g^{*}=0$。

3 实验仿真与分析

实验数据集使用 KDDCUP1999[8] 中的2个10%独立子集作为训练样本集和测试数据集,在进行训练和样本集预处理时,将训练样本集随机地分为3组,其中1组为训练样本,其余2组为增量数据集。具体的数据集如表1所列。

表1 实验数据集表

攻击类别	训练样本集			测试样本集		
	样本数目	正常样本	攻击样本	样本数目	正常样本	攻击样本
Dos	9000	5863	3137	3000	1956	1046
Probe	4500	3671	829	1500	1224	276
R2L	4500	3815	685	1500	1276	224
U2R	2250	2229	21	750	723	27

算法实现以 Matlab7.1 和 libsvmmat2.911 工具包为基础,其中 SVM 的核函数采用径向基函数 RBF,核参数及惩罚系数利用交叉验证参数的方法赋值,聚类数目 $k=\left\lfloor\frac{n}{2}\right\rfloor$,$n$ 为聚类样本数目。通过和文献[5]的 B-ISVM 算法进行对比,对该算法的检测率、误报率和训练时间进行了验证,具体如图1—图4所示。

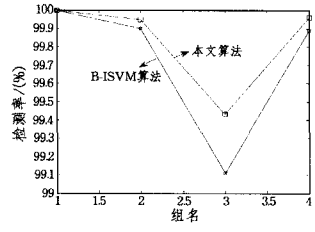

图1 两种算法的检测率对比

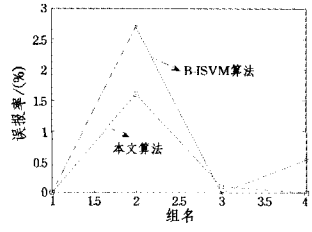

图2 两种算法的误报率对比

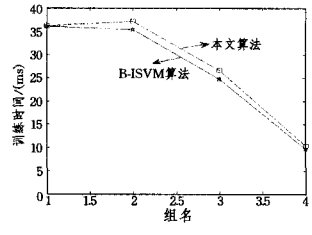

图3 两种算法的训练时间对比

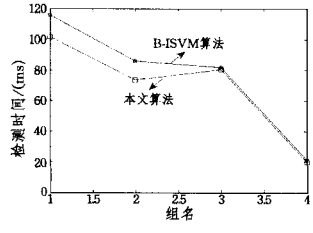

图4 两种算法的训练时间对比

通过对比可以看出,本文算法以较小的训练时间为代价,使得整体性能要优于 B-ISVM 算法,以 Probe 攻击为例,以4.8%的训练时间为代价,在保证检测率为100%的前提下使得误报率降低了40.74%,检测时间减少了14.55%。

结束语 通过对基于距离加权模板约简的 K 近邻聚类分析算法和基于属性信息熵分类超平面构造进行研究,提出了一种基于距离加权模板约简和属性信息熵的增量 SVM 入侵检测算法。该算法对正负样本集进行聚类分析,通过邻界子聚类矩阵确定邻界区,并利用属性信息熵对噪声样本点和过拟合样本点进行剔除,进而构造分类超平面,并以样本离散度和支持向量提取为方式,基于 KKT 条件进行增量学习,从而获取最终的最优超平面,实现 SVM 分类器构造。实验仿真可以看到,该算法具有较好的检测率和检测速度,且分类效果好,误报率低。

参考文献

[1] 黄双福,陈贤富.基于改进 SVM 主动学习算法的入侵检测[J].微电子学与计算机,2010,27(3):75-77

[2] 李汉彪,刘渊.一种 SVM 入侵检测的融合新策略[J].计算机工程与应用,2012,48(4):87-90

(下转第86页)

4 实验结果及其分析

4.1 数据集及度量标准

实验采用 MovieLens 站点(http://movielens.umn.edu)提供的数据集。该站点是一个基于 Web 的研究型推荐系统,用于接收用户对电影的评分并提供相应的电影推荐列表。该数据集中每个用户至少对 20 部电影进行了评分。整个实验数据集根据实验需要进一步地划分为训练集和测试集。其中训练集占总数的 80%,测试集占 20%。

采用平均绝对偏差 MAE 作为度量标准。平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差来度量预测的准确性,MAE 越小,推荐质量越高。设预测的用户评分集合表示为$\{p_1,p_2,\cdots,p_N\}$,对应的实际用户评分集合为$\{q_1,q_2,\cdots,q_N\}$,则平均绝对偏差 MAE[11]定义为:

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{9}$$

4.2 实验结果

实验中首先根据用户情景信息对所有用户进行模糊聚类,选取最佳聚类阈值 $\lambda=0.7$,并以相关相似性作为寻找最近邻的相似度计算方法。将本文算法与传统的协同过滤推荐算法以及文献[12]提出的一种改进的协同过滤推荐算法进行对比,在计算 MAE 时,邻居个数从 5 个增至 30 个,间隔为 5。实验结果对比如图 2 所示。

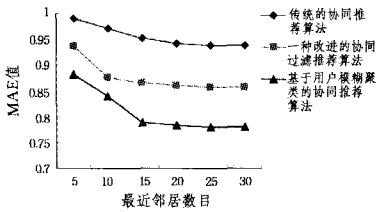

图 2 实验结果图

从图 2 可以看出,随着目标用户的最近邻居数目的增加,上述 3 种算法的 MAE 值都呈现逐渐下降的趋势。但是在各种实验条件下,本文提出的改进算法均具有最小的 MAE 值。由此可知,本文提出的基于用户模糊聚类的协同过滤推荐算法可以有效地缓解评分矩阵稀疏的问题,提高推荐系统的推荐质量。

4.3 实验结果分析

传统的协同过滤推荐算法是在整个用户空间上根据所有用户对项目的评分来计算用户间的相似性,不仅计算量大,而且将所有用户对项目评分的重要性视为同等的,忽略了不同特征用户之间评分的差异性,这无疑会降低系统推荐的准确性。一种改进的协同过滤推荐算法只考虑了用户评分对项目相似性的影响,而没有充分利用已有的用户数据和领域知识,特别是在评价数据稀疏的情况下,造成项目相似性计算不准确的问题,严重影响了推荐的精确度。而本文充分考虑了用户自身信息对评分的影响,协同过滤是在具有相似情景的同类用户中进行的,同类用户更容易产生相似的兴趣爱好,同时,在协同过滤前对稀疏评分矩阵进行了填充,这大大提高了近邻用户选取的准确性。实验结果表明,本文提出的算法有效地提高了推荐系统的推荐质量。

结束语 随着推荐系统规模越来越大,用户数目和项目数目急剧增加,推荐系统的实时性要求越来越难以满足。本文提出的算法引入用户情景因素,通过模糊聚类技术对稀疏用户-项目评分矩阵进行降维,然后对降维后的评分矩阵进行填充,最后利用协同过滤在线进行推荐。把聚类分析用于协同过滤推荐中,既降低了用户空间的维度,使得搜寻最近邻居用户的范围缩小,又能够提高协同过滤算法的可扩展性和效率。实验结果表明,基于用户模糊聚类的协同过滤算法有效地解决了数据稀疏性问题,提高了推荐的准确度。

参 考 文 献

[1] 李聪,梁昌勇,马丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展,2008,45(9):1532-1538

[2] 张光卫,康建初,李鹤松,等. 面向场景的协同过滤推荐算法[J]. 系统仿真学报,2006,18(z2):595-601

[3] 郑先荣,曹先彬. 线性逐步遗忘协同过滤算法的研究[J]. 计算机工程,2007,33(6):72-74

[4] Balabanovic M, Shoham Y. Fab: Content-based Collaborative Recommendation[J]. Communication of the Association for Computing Machinery,1997,40(3):66-72

[5] 秦围,杜小勇. 基于用户层次信息的协同推荐算法[J]. 计算机科学,2004,31(10):138-140

[6] 杨虎,刘琼荪,钟波. 数理统计[M]. 北京:高等教育出版社,2004:190-199

[7] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). 1998:43-52

[8] Sarwar B, Karypis G, Konstan J, et al. Item-Based collaborative filtering recommendation algorithms[C]// Proceedings of the 10th International World Wide Web Conference. 2001:285-295

[9] 罗辛,欧阳元新,熊璋,等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报,2010,33(8):1437-1445

[10] 曾春,邢春晓,周立柱. 个性化服务技术综述[J]. 软件学报,2002,13(10):1952-1961

[11] 张丙奇. 基于领域知识的个性化推荐算法研究[J]. 计算机工程,2005,31(21):7-9

[12] 刘旭东,陈德人,王惠敏. 一种改进的协同过滤推荐算法[J]. 武汉理工大学学报:信息与管理工程版,2010,32(4):550-553

(上接第 78 页)

[3] 杨金福,宋敏,李明爱. 一种新的基于距离加权的模板约简 K 近邻算法[J]. 电子与信息学报,2011,33(10):2378-2383

[4] 童先群,周忠眉. 基于属性值信息熵的 KNN 改进算法[J]. 计算机工程与应用,2010,46(3):115-117

[5] 牟琦,陈艺坤,毕孝儒. 一种基于快速增量 SVM 的入侵检测方法[J]. 计算机工程,2012,38(12):92-94

[6] Fayed H A, Atiya A F. A novel template reduction approach for the k-nearest neighbor method[J]. IEEE Transactions on Neural Networks,2009,20(5):890-896

[7] 夏书银,王越,张权. 核空间结合样本中心角度的支持向量机增量算法[J]. 计算机应用与软件,2012,4:121-124

[8] KDD99Cupdataset[EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,2011-01-14

[9] 肖敏,柴蓉,杨富平,等. 基于可拓集的入侵检测模型[J]. 重庆邮电大学学报:自然科学版,2010,22(3):345-349