文章编号:1671-4598(2017)08-0214-04 DOI:10.16526/j.cnki.11—4762/tp.2017.08.055 中图分类号:TP393 文献标识码:A

基于特征选择的网络入侵检测模型研究

李 文
(广东科贸职业学院 信息工程系,广州 510640)

摘要:为了有效从收集的恶意数据中选择特征去分析,保障网络系统的安全与稳定,需要进行网络入侵检测模型研究;但目前方法是采用遗传算法找出网络入侵的特征子集,再利用粒子群算法进行进一步选择,找出最优的特征子集,最后利用极限学习机对网络入侵进行分类,但该方法准确性较低;为此,提出一种基于特征选择的网络入侵检测模型研究方法;该方法首先以增强寻优性能为目标对网络入侵检测进行特征选择,结合分析出的特征选择利用特征属性的 Fisher 比构造出特征子集的评价函数,然后结合计算出的特征子集评价函数进行支持向量机完成对基于特征选择的网络入侵检测模型研究方法;仿真实验表明,利用支持向量机对网络入侵进行检测能有效地提高入侵检测的速度以及入侵检测的准确性。

关键词:特征选择;网络入侵;Fisher 比;支持向量机

Network Intrusion Model Based on Feature Selection Research

Li Wen
(Department of information Engineering, Guangdong Polytechnic of Science and Trade, Guangzhou 510640, China)

Abstract: In order to effectively extract features from the malicious data collected to analyze, security network system security and stability, the need for network intrusion detection model is studied. But the current approach is to use genetic algorithm to find out the characteristics of the network intrusion subset of recycled for further selection of particle swarm optimization (pso), find out the optimal feature subset, finally using extreme learning machine classifying network intrusion, but this method has the problem of accuracy is low. Therefore, proposes a network intrusion detection methods based on feature selection. This method firstly in order to enhance optimal performance as the goal to feature selection of network intrusion detection, combined with analysis of characteristics of feature selection using the attributes of the Fisher than feature subset evaluation function is constructed, and combining with the feature subset of calculated results of evaluation function for support vector machine (SVM) to network intrusion detection based on feature selection methods. Simulation experiments show that support vector machine (SVM) is used to analyse the network intrusion detection can effectively improve the accuracy of the speed of intrusion detection and intrusion detection.

Keywords: mobile application platform; Network security; Assessment

0 引言

随着互联网技术应用的日渐广泛,互联网络的安全性以及可靠性越来越受到人们的关注[1]。互联网络平台是一个双边平台,具有共享性与开放性的特点,由于互联网络的开放性,加上入侵手段的多样化[2],网络的恶意入侵越来越频繁。在这种情况下,如何提高网络入侵的检测率和检测速度,保证互联网络的正常通信与数据运输安全成为了网络管理领域中急需解决的主要问题[3]。对于现有的网络恶意入侵的检测方法有很多,这是在不断更新、不断发展的网络主动式的自我防御策略技术,利用网络相互之间发生联系时的动态特征来准确描述此时网络是否受到了入侵,这项技术在当前网络安全保护技术的发展中起着至关重要的作用[4]。随着网络复杂度的增高以及网络需求速度的提升,恶意入侵行为日益增加,这是出现的明显问题为不能对网络传输的数据进行实时处理,网络入侵检测的复杂混乱特征的提纯以及对入侵过程信息处理分析导致了对入侵检测过程复杂度的增高,致使检测时间加长[5]。而有效地对特征选择的网络入侵进行检测是解决上述问题的有效途径。已引起了该领域专家和学者的关注与重视,由于网络入侵检测具有广泛的发展空间,因此,成为了计算机网络检测研究的核心,具有较大的发展潜力[6]。

近年来取得了一定的成果,裴恩斯提出了网络入侵检测系统的创建模型,根据该模型对网络恶意入侵行为进行有效快速的检测,利用在入侵过程主动记录下的数据信息来构建关联系统框架,通过对该框架的变化程度来对网络入侵行为进行监测[7]。郎恩提出了基于神经网络的检测入侵系统,该系统利用图论对网络执行检测入侵功能,解决大多数入侵识别检测系统的稳定性不够的问题,利用数据信息统计表来对不同种类入侵攻击行为下存在的联系以及区别,创建不同攻击类型之间关系模型。王宇航提出了基于数据挖掘框架自适应的入侵检测方法,通过审计程序对网络会话连接的特征集进行提取,然后利用数据挖掘算法在数据特征集上表达入侵行为模式,采用这种模式对入侵进行指导。文献[8]提出一种基于遗传算法选择特征的网络入侵检测方法,通过遗传算法找出网络入侵的特征子集,再利用粒子群算法进行进一步选择,找出最优的特征子集,最后利用极限学习机对网络入侵进行分类,但该方法存在准确性较低的问题。文献[9]提出一种参数优化的特征选择网络入侵检测方法。该方法首先将检测的准确率作为问题优化

收稿日期:2017-04-15; **修回日期**:2017-04-26。
基金项目:医学院校课程考试与学业评价管理通用系统的改革与研究(桂教科研[2003]22号)。
作者简介:李 文(1963-),男,广西钦州人,硕士研究生,副教授,主要从事计算机网络应用、软件应用、网络安全、大数据、云安全、软件开发等方向的研究。

的主要目标函数，网络特征与参数作为约束条件建立检测模型，通过对检测模型进行求解，找出最优的特征子集和最优参数，但该方法存在过程较为复杂的问题。文献［10］提出一种特征优化耦合的网络入侵检测模型。首先通过径向函数将网络特征映射到高维空间内对此进行计算，建立网络特征和网络入侵分类器间的联系，在特征提取阶段解决了分类器参数的设计问题，建立网络入侵的检测模型，但该方法存在检测速度较慢的问题。

针对上述问题，提出一种基于特征选择的网络入侵检测模型研究方法。该方法首先以增强寻优性能为目标对网络入侵检测进行特征选择，结合分析出的特征选择利用特征属性的 Fisher 比构造出特征子集的评价函数，然后结合计算出的特征子集评价函数结果进行支持向量机完成对基于特征选择的网络入侵检测模型研究方法。仿真实验表明，利用支持向量机对网络入侵进行检测能有效地提高入侵检测的速度以及入侵检测的准确性。

1　基于特征选择的网络入侵检测模型研究

首先以增强寻优性能为目标对网络入侵检测进行特征选择，结合分析出的特征选择利用特征属性的 Fisher 比构造出特征子集的评价函数，然后结合计算出的特征子集评价函数结果进行支持向量机完成对基于特征选择的网络入侵检测研究方法。具体步骤如下：

1.1　网络入侵检测特征选择

网络入侵检测的特征可用二进制字符来表示：$S=\{s_1,s_2,\cdots,s_n\}$，$s_i\in\{0,1\}$，$i=1,2,\cdots,m$，其中“1”代表较优特征，且被选中，反之，“0”代表没有被选择上的特征，m 代表网络入侵数据特征的整体维数，因此特征选择的数学模型为：

$$\max G(S)\quad s.t\begin{cases}S=\{s_1,s_1,\cdots,s_n\}\\ s_i\in\{0,1\}\\ i=1,2,\cdots,n\end{cases}\tag{1}$$

由该公式可以推断出，在网络受到入侵的情况下，对满足约束的最优特征子集的寻找是该问题中较为典型的组合优化问题。对网络入侵中特征求取过程无法实现对特征的选择，所以需要先对网络入侵数据特征进行编码。

特征选择的目标是选择较少的特征，获取更高的网络入侵检测的检测率，由此适应度函数的定义为：

$$f=\omega_a\times Acc+\omega_f\Big(\sum_{i=1}^{N_f}f_i\Big)^{-1}\tag{2}$$

公式（2）中，ω_a 为特征数量的权重，本文的取值是 0.6，N_f 为特征的总数，Acc 为验证集网络入侵检测的正确率，ω_f 是权重，本文的取值是 0.4，f_i 为特征选择的状态，即：

$$f_i=\begin{cases}1,\text{第 } i \text{ 个特征被选中}\\ 0,\text{第 } i \text{ 个特征未被选中}\end{cases}\tag{3}$$

对于 d 维特征空间中由 n 个样本构成的网络数据集 $X=\{x_1,x_2,\cdots,x_n\}$，$x_i(i=1,2,\cdots,n)\in R^d$，可划分为 c 个类：C_1，$C_2,\cdots,C_c$，每类中包括 n_i 个样本，$\sum_{i=1}^{c}n_i=n$。

假设第 k 维特征在网络样本集上类间的离散度 $S_b^{(k)}$ 与类内的离散度 $S_w^{(k)}$ 的比值：$S_b^{(k)}/S_w^{(k)}$ 为网络特征的 Fisher 比。

$$S_b^{(k)}=\sum_{j=1}^{c}\frac{n_j}{n}(m_j^{(k)}-m^{(k)})^2\tag{4}$$

$$S_w^{(k)}=\sum_{j=1}^{c}\Big(\frac{1}{n_j}\sum_{x\in Cj}(x^{(k)}-m_j^{(k)})^2\Big)\tag{5}$$

上述公式中，$x^{(k)}$ 表示网络数据样本 x 的第 k 维特征空间，$m_j^{(k)}$ 表示第 j 类网络数据样本的第 k 维特征的均值，$m^{(k)}$ 表示所有网络数据样本的第 k 维特征的均值，$S_b^{(k)}$ 和 $S_w^{(k)}$ 分别表示不同类数据样本的距离和同类数据样本的距离。

假设由 k 个特征构成的特征子集在数据样本集中的类间离散度 $S_b^{(Rk)}$ 与类内离散度 $S_w^{(Rk)}$ 的比值：$S_b^{(Rk)}/S_w^{(Rk)}$ 为网络特征子集的 Fisher 比。

$$S_b^{(Rk)}=\sum_{i=1}^{k}\sum_{j=1}^{c}\frac{n_j}{n}(m_j^{(i)}-m^{(i)})^2\tag{6}$$

$$S_w^{(Rk)}=\sum_{i=1}^{k}\sum_{j=1}^{c}\Big(\frac{1}{n_j}\sum_{x\in Cj}(x^{(i)}-m_j^{(i)})^2\Big)\tag{7}$$

为对特征选择进行简化计算，将网络入侵检测的数据样本分为两类：正常数据类与入侵数据类，称为正类样本和负类样本，将网络入侵检测问题简化为二分类问题。对上述的网络样本数据集 $X=\{x_1,x_2,\cdots,x_n\}$，将 X 正类数据样本集记为 X_1，负类数据样本集 X_2，n_1 为正类样本数，n_2 为负类数据样本数，依据公式（6），公式（7）得：

$$S_b^{(Rk)}=\sum_{i=1}^{k}\Big(\frac{n_1}{n}(m_j^{(i)}-m^{(i)})^2+\frac{n_2}{n}(m_2^{(i)}-m^{(i)})^2\Big)\tag{8}$$

$$S_w^{(Rk)}=\sum_{i=1}^{k}\Big(\frac{1}{n_1}\sum_{x\in X_1}(x^{(i)}-m_1{}^{(i)})^2+\frac{1}{n_2}\sum_{x\in X_2}(x^{(i)}-m_1^{(i)})^2\Big)\tag{9}$$

Fisher 比可以反映出网络入侵特征检测对数据分类的影响以及作用，该比值可大可小，比值越大，那么相对应的特征子集的分类能力就越强。因此，特征子集评价函数为：

$$F=\sum_{i=1}^{k}\Big(\frac{n_1}{n}(m_1^{(i)}-m^{(i)})^2+\frac{n_1}{n}(m_2^{(i)}-m^{(i)})^2\Big)\Big/\sum_{i=1}^{k}\Big(\frac{1}{n_1}\sum_{x\in X_1}(x^{(i)}-m_1^{(i)})^2+\frac{1}{n_1}\sum_{x\in X_2}(m^{(i)}-m_2^{(i)})^2\Big)\tag{10}$$

1.2　基于支持向量机的入侵检测模型

结合上述分析出的网络特征选择为基础，利用支持向量机对选择出的特征进行分类，支持向量机为训练数据集的子集，定义了超平面，把数据集分成 2 类。对于不能分成 2 类的情况，可把数据映射到高维特征空间中进行解决。支持向量机为凸优化问题，局部的最优解就是全局最优解。

假设有 2 类线性可分的数据样本集合：(x_i,y_i)，$i=1,2,\cdots,n$，$x_i\in R^d$，$y_i\in\{+1,-1\}$，满足条件：

$$y_i[(\omega\cdot x_i)+b]-1\geqslant 0,i=1,2,\cdots,n\tag{11}$$

使 $\frac{1}{2}\|\omega\|^2$ 最小的分类面为最优的分类面，利用 Lagrange 优化方法可把网络入侵最优分类面的求解问题转换为凸二次规划寻优的对偶问题：

$$\max\sum_{i=1}^{n}a_1-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_j(x_i,x_y)\tag{12}$$

其中：

$$a_1\geqslant 0,i=1,2,\cdots,n\tag{13}$$

约束条件为：

$$\sum_{i=1}^{n}a_1y_i=0\tag{14}$$

公式（14）中，a_1 表示 Lagrange 乘子，为二次函数寻优的

问题，存在唯一的解。可证明，在方程解中存在部位 0 的 a_1，且不唯一，这些 a_1 所对应的向量即为是支持向量机。根据以上的求解，得出最优分类面函数为：

$$f(x) = sga\{(\omega \cdot x) + b\} = \mathrm{sgn}\left\{\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\right\} \tag{15}$$

假设最优分类面不能把 2 类点分开时，可通过引入松弛因子 $\xi(\xi \geqslant 0)$，这种情况下允许错分数据样本的存在。此时：

$$\varphi(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \tag{16}$$

公式（16）中，C 表示惩罚因子，可得出广义的最优分类面。广义最优分类面的对偶问题与线性分类情况完全相同，只是把公式（13）改为：

$$0 \leqslant a_i \leqslant C, I = 1, 2, \cdots, n \tag{17}$$

对于分线性分类问题，可把相关关联数据组进行映射处理，映射到高维空间后，进而实现关联特征的线性分类来解决问题。此时特征相对应的分类函数为：

$$f(x) = \mathrm{sgn}\left\{\sum_{i=1}^{n} a_i^* y_i K(x_i, x) + b^*\right\} \tag{18}$$

公式（18）中，K 表示函数，满足条件的函数包括：1）多项式函数：$K(x_i, x) = [(x \cdot x_i) + 1]^q$；2）径向函数：$K(x_i, x) = \exp\left\{-\frac{|x \cdot x_i|^2}{\sigma^2}\right\}$；3）Sigmoid 函数：$K(x_1, x) = \tanh[v(x \cdot x_i) + a]$，由此完成对基于特征选择的网络入侵检测模型研究。

2 实验结果与分析

为了证明基于选择特征的网络入侵检测模型研究方法的有效性，需要进行一次仿真实验。选择 KDD2016 数据集作为仿真对象，数据集包括拒绝攻击（DoS）、未授权远程访问（Probe）、扫描与探测（R2L）以及对本地用户非法访问（R2R）4 种攻击方式，其余数据为正常数据。实验采用 Intel 奔腾 43.0CPU、内存为 2 G 的计算机上进行，在 Matlab2016 上进行编程实现。

为了使检测结果更具有说服力，在相同的实验下与相同的数据集进行对比实验，在实验中主要对文献［8］给出的遗传算法和文献［9］给出的方法和本文方法进行特征选择时的性能差异。在测试实验中选取 100 次运行的平均值作为性能差异对比结果。

利用下述公式计算检测率：

$$WA = NB_{ce}/NB_{ei} \times 100\% \tag{19}$$

利用下述公式计算漏检率：

$$CN = BA_{sd}/BA_{fg} \times 100\% \tag{20}$$

其中：NB_{ce} 表示网络入侵数据次数、NB_{ei} 表示网络入侵异常次数、BA_{sd} 表示网络入侵漏检次数、BA_{fg} 表示网络入侵全部次数。

表 1 不同方法性能对比

攻击类型	检测时间			检测的准确率/%		
	本文方法	文献[8]	文献[9]	本文方法	文献[8]	文献[9]
DoS	0.231	0.331	0.331	97.95	96.2	92.1
Probe	0.190	0.416	0.416	98.1	93.25	94.32
R2L	0.212	0.551	0.551	98.2	95.62	97.2
R2R	0.221	0.410	0.462	99.2	98.53	96.3

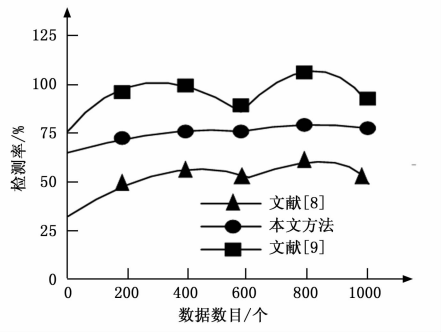

图 1 不同方法的检测率（%）

从表 1 和图 1 可看出，本文提出的基于特征选择的网络入侵检测方法与文献［8］中给出的遗传算法和文献［9］中给出方法相比较，在检测时间方面，本文方法的时间最少、表现最好，网络入侵的检测率明显高于文献［8］和文献［9］两种方法，区别很明显，能看出本文的方法能更有效地对网络数据进行精简，在检测的时间以及检测的准确性方面的表现明显优于文献［8］和文献［9］中的方法，能有效地解决网络入侵检测特征选择存在的问题，保证较高的准确率。

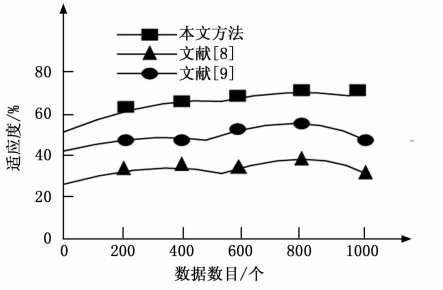

图 2 不同方法的网络入侵适应度对比

由图 2 可知，文献［8］中给出的遗传算法的适应度较差，虽然浮动很均匀，但随着数据数目的增加，适应度越来越低，文献［9］给出的方法虽然比文献［8］的适应度会高一些，但总体来说可行性较差，本文所提方法的适应度较强，随着数据数目的增加适应度也越来越高，虽然也略有波动，但和文献［8］、文献［9］相比，本文方法的网络入侵检测的适应度较高。

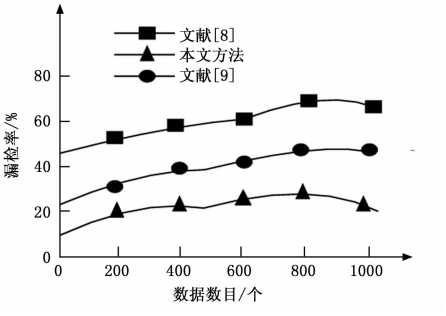

图 3 不同方法的数据漏检率（%）对比

由图 3 可看出文献［8］给出方法的漏检率随着数据数目的增加，漏检率越来越高，那么该方法检测的准确性就会降

低，文献［9］给出方法的漏检率相对于文献［8］较低，但随着数据数目的增加，漏检率也在逐渐的提高，由此看出文献［8］和文献［9］给出的方法可行性较低，而本文方法随着数据数目的增加漏检率越来越低，由此可说明本文方法的准确性较高。

仿真实验表明，本文所提方法能有效地提高对网络入侵行为进行检测，并且保证了入侵检测的准确率。

3　结论

采用遗传算法找出网络入侵的特征子集，再利用粒子群算法进行进一步选择，找出最优的特征子集，最后利用极限学习机对网络入侵进行分类，但该方法准确性较低。为此，提出一种基于征选择的网络入侵检测模型研究方法。并通过实验证明，本文所提方法能有效地提高特征选择的网络入侵检测模型的准确性，具有广泛的实用价值。

参考文献：

［1］唐成华，刘鹏程，汤申生，等. 基于特征选择的模糊聚类异常入侵行为检测［J］. 计算机研究与发展，2015，52（3）：718－728.

［2］张　拓，王建平. 基于 CQPSO－LSSVM 的网络入侵检测模型［J］. 计算机工程与应用，2015，51（2）：113－116.

［3］刘白璐，杨雅辉，沈晴霓. 一种基于遗传算法的入侵早期特征选择方法［J］. 小型微型计算机系统，2015，36（1）：111－115.

［4］黄春虎，努尔布力，解男男，等. 基于 Re－FCBF 的入侵特征选择算法研究［J］. 激光杂志，2016，37（1）：103－107.

［5］唐　喆，曹旭东. 网页分类中特征选择方法的研究［J］. 电子设计工程，2016，24（5）：120－122.

［6］武小年，彭小金，杨宇洋，等. 入侵检测中基于 SVM 的两级特征选择方法［J］. 通信学报，2015，36（4）：19－26.

［7］姜　宏，陈庶樵，扈红超，等. 基于 GAIG 特征选择算法的轻量化 DDoS 攻击检测方法［J］. 计算机应用研究，2016，33（2）：502－506.

［8］黄　亮，吴　帅，谭国律，等. 基于 EPSO－RVM 的网络入侵检测模型［J］. 计算机工程与应用，2015，51（3）：85－88.

［9］梁　辰，李成海，周来恩. PCA－BP 神经网络入侵检测方法［J］. 空军工程大学学报：自然科学版，2016，17（6）：93－98.

［10］余文利，余建军，方建文. 一种新的基于 KPCA 和改进 ε－SVM 的入侵检测模型［J］. 计算机工程与应用，2015，51（11）：93－98.

（上接第 205 页）

4.2　关键功能测试

手掌姿态传感器数据获取测试，右手佩戴手套，手心朝下，小臂平行于地面，以手肘为圆心旋转小臂大约 120°，不断往复，手指传感器数据获取测试，右手佩戴手套，手掌朝下，握拳、张开手指，不断重复此动作，在虚拟示波器得到各传感器的数据波形，可以看到 5 个手指的弯曲的传感器的值波动较大。

将手势数据通过蓝牙传输给中继控制端，中继控制端根据手势短时间上的匹配来进行不同语义上的解析，中继设备从而实现对电机和舵机的间接控制，佩戴人员实现对设备的直接控制，大致控制手势调整步骤如下：

手平移到被控设备上（电机、舵机）进入设备待选模式；单手抱拳，表示选择该设备，进入手势控制模式；如果选择的是电机，此时手掌前后倾角大小直接用于控制电机转速和方向；如果选中的是舵机，手掌作用倾角大小直接用于控制舵机的打角；手掌平移回到没有设备的中间地方，单手抱拳，解除设备的控制状态，设备维持在上一控制量上，回到步骤（1）的设备选择状态；如果发现突发情况，可以按下中继板上的 TSI 按键可以紧急制动，即紧急停止所有设备。

4.3　可靠性测试

佩戴满电状态的手套使用 3 小时，重复做一组连续的动作，将姿态编码发送到上位机上，与手势的序列做对比，获得的 431 组数据，其中没有匹配上的数据位 73 组，准确率达到 82%，此过程中系统稳定工作，没有发送程序跑飞或硬件错误状态，静电、电磁干扰等环境的适应能力强。

5　结论

本设计采用分层设计思想，将手势系统分解为手套采集部分、中继混合控制部分、上位机分析处理部分和网络云端部分，设计出具体的人机交互融合方案，让可穿戴设备能够实现更加智能的数据交互。该手势交互手套可应用在智能家居领域、工业控制方向、远程操作工作、手语翻译研究等多种领域，在操作、维护等方面充分考虑了人性化设计，具有很强的可行性，关键功能测试与性能验证达到预期标准。本设计充分考虑了未来功能升级、规模扩展潜在需求，不同层次的部分均可加以扩展和应用。该手势手套具备很强的新颖性和实用性，有良好的推广价值。

参考文献：

［1］孟祥旭，李学庆. 人机交互技术－原理与应用［M］. 北京：清华大学出版社，2005.

［2］Grimes G J. Digital data entry glove interface device［P］. Technical Report US Patent，2011：1561－1566.

［3］Lee C，Xu Y. Online interactive learning of gestures interfaces［A］. Proceeding of IEEE International Conference on Robotics and Automation［C］. 2013：156－158.

［4］董士海. 人机交互进展及面临的挑战［J］. 计算机辅助设计与图形学报，2004（10）：20－21.

［5］张海涛，阎贵平. 加速度传感器的原理及分析［J］. 电子设计技术，2003（21）：21－23.

［6］付梦印，邓志红，张继伟. Kalman 滤波理论及其在导航系统中的应用［M］. 北京：科学出版社，2003.

［7］修国浩. 基于 WD/HMM 的语音识别算法研究［D］. 秦皇岛：燕山大学，2004.

［8］江　浩，褚衍东，郭丽峰. 曲线形态相似性的定义与度量［J］. 云南民族大学学报，2009（5）：54－56.

［9］邹　波，张　华. 多传感器信息融合的改进扩展卡尔曼滤波定姿［J］. 计算机应用研究，2012（8）：8－10.

［10］敬　喜. 卡尔曼滤波器及其应用基础［M］. 北京：国防工业出版社，2014.

［11］郝立果. 基于加速度传感器的运动信息采集和应用研究［D］. 天津：天津大学，2009.

［12］曹　赟，周　宇，徐寅林. 加速度传感器在步态信号采集系统中的应用［J］. 信息化研究，2009（7）：30－31.