基于聚类分析的软件胎记特征选择

罗养霞[1,2],房鼎益[2]
(1.西安财经学院信息学院,陕西西安 710100;2.西北大学信息科学与技术学院,陕西西安 710127)

摘 要: 软件胎记选择关系着软件的识别率.本文应用约束聚类分析软件特征,基于互信息度量特征的类内和类间距离,以同类和异类软件特征构建信息增益函数和惩罚函数,选择出具有高的类区分信息和最小冗余的软件胎记特征.通过分析和比较表明该算法为软件胎记特征的选择和优化提供了一种有效途径.

关键词: 胎记特征选择;聚类分析;信息度量;关联系数
中图分类号: TP391.1 **文献标识码:** A **文章编号:** 0372-2112 (2013) 12-2334-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.12.003

Feature Selection for Software Birthmark Based on Cluster Analysis

LUO Yang-xia[1,2], FANG Ding-yi[2]
(1. *School of Information, Xi'an University of Finance and Economics, Xi'an, Shaanxi* 710100, *China*;
2. *School of Information Science and Technology, Northwest University, Xi'an, Shaanxi* 710127, *China*)

Abstract: The feature selection for software birthmark has a direct bearing on software recognition rate. We apply constrained clustering to analyze software features. The within-and between-class distances of features are measured based on mutual information. Information gain and penalty functions are constructed using homogeneous and heterogeneous software features respectively. Then the software birthmark features with high class distinction and minimum redundancy are selected. It is shown the algorithm provide an effective approach for software birthmark feature selection and optimization by analysis and comparison.

Key words: birthmark feature selection; clustering analysis; information measurement; correlation coefficient

1 引言

软件胎记研究是提取软件中的不变特征(或关键特征),以达到对该软件或该软件家族的识别,检测软件盗版或阻止恶意软件的破坏.若两个软件的胎记特征之间具有较大的相似度,则其中一个软件很可能是另外一个软件的盗版,或者与另外一个软件属于同一家族,可见,软件特征的分析和选择关系着软件识别的鲁棒性和可信性.

传统的软件特征获取技术有:(1)基于静态的特征提取[1],得到的特征对于软件的加密、变形、多态攻击已不能满足[2];有些基于 n-gram 分段式特征选择[3,4],只考虑了程序的语法结构,而没有结合语义分析;(2)为了深层分析动态特征,通常结合动态程序切片[5]对特征进行过滤,但这种选择方式仍含有较大的特征冗余,检测效率低;结合语义分析的行为特征[6],需要代码在虚拟机中虚拟执行,获得相应的执行行为和路径,检测开销较大,构建的特征库庞大.Tamada H 等研究[7]指出要选择难以被混淆的特征,如具有原子性,全局性,不可伪造性.在 Unix 系统中,read 就不适合作 API 胎记,由于它不具有原子性,可被分成两个或多个调用;gettimeofday()和 getpid()由于可能被任意增减也不适合作胎记特征.目前,虽然特征选择方法有很多,但如何针对软件胎记特征这一特定问题给出有效的方法,仍需要进一步研究解决.

论文提出基于聚类分析筛选软件特征,改进传统的分段或切片式选择.研究软件变化时特征的变化规律,研究基于层次聚类中距离的度量方法,研究互信息度量函数的构成,最终,按类间相似度最小,类内相似度最大的原则,选择出分类信息较高的特征集合为胎记(或不变)特征.该研究的优点是:(1)聚类动态分析特征,考虑了特征间的相关性,力求组成的特征集合分类信息量

收稿日期:2012-10-21;修回日期:2013-02-20;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61070176,No.61170218,No.61272461);国家级大学生创新项目(No.201311560008);陕西省教育厅科研计划项目(No.2013JK1200);陕西省自然科学基础研究(No.2012JM8034);陕西省教育厅产业化项目(No.2011JG06)

最大且冗余度最小;2)以互信息度量距离,构建增益函数和惩罚函数,既选择同类中抗等价语义变化攻击的不变特征,又剔除不同类中的普遍特征,保证胎记特征的抗攻击性和唯一性.通过实验和分析表明该研究在软件胎记特征选择和过滤方面具有较好的效果,实用性更强.

2　相关知识

2.1　等价语义变换

软件特征(Software Features,SF)来自于程序相关的信息,包括程序的类继承关系,字节码,操作码,API 调用频率,线程执行序列,控制流,数据流,软件时空结构等一系列的信息.将软件 P 按照软件黑盒思想,即是将软件 P 看成一个进行输入 I 后产生输出 O 的黑盒.将软件特征分成软件输入特征(Input Aoftware Feature,ISF),软件输出特征(Output Software Feature,OSF),软件自身特征(Self Software Feature,SESF).可以将 SF 看成 ISF,OSF 和 SESF 的并集,即为 SF←ISF∪OSF∪SESF.

定义 1　等价语义变换(Semantics-Preserving Transformation,SPT),软件 P_i 是软件 P 经过变换方法得到的 $P_i = \mathrm{SPT}(P)$,使得 $\mathrm{ISF}_P = \mathrm{ISF}_{Pi}$ 且 $\mathrm{OSF}_P = \mathrm{OSF}_{Pi}$,但 $\mathrm{SESF}_P \neq \mathrm{SESF}_{Pi}$,则称此变换方法是等价语义变换.

由于软件具有等价语义变换能力[8],从恶意软件角度看,也称为对软件的攻击能力.等价语义变换会引起软件特征的变化,不易发生变化的特征称为胎记特征.对同一软件 P_o 进行不同等价语义如:混淆,优化,压缩等变化后,即 $P_i = \mathrm{SPT}(P_o)$,充分得到 P_o 的多样性表现形式 $P_i = \{p_1, p_2, \cdots, p_i\}$,研究中定义 P_i 是原软件 P_o 的等价软件(或同类软件).参照程序为 $Q_j = \{q_1, q_2, \cdots, q_j\}$,是与 P_o 具有相同功能,但属于不同版本的异类软件.由鲁棒性和可信度定义[9],可知胎记特征要求在等价语义变化下,需有较高的鲁棒性,不易受各种变换类攻击的影响,即变化前后所选择的胎记特征应在同类软件 P_i 中相似度越高越好,但与不同类 Q_j(参照程序)中相似度越低越好,以保持胎记特征的唯一性和可信性.

2.2　特征选择

针对实际目标要求的不同,特征选择算法关注的侧重点也不同.设数据特征集 $F = \{f_1, f_2, \cdots, f_n\}$,在原始特征 F 中寻找一个子集 $B \subseteq F$,特征子集的评价函数为 $J(X): 2^F \rightarrow [0,1]$,其中 $J(X)$ 值越大,表示特征子集 X 所含信息量更多,0 或 1 表示特征的选择与否,所有取值为 1 的特征为最后被选的子集.特征选择算法按侧重点不同可分为三种类型[10]:

(1)从特征集 F 中找到一个特征子集 X,使得 $J(X)$最大;

(2)给定阈值 J_0,从 F 中找到一个最小子集 X,使得 $J(X) > J_0$;

(3)从 F 中找到一个子集 X,使得 $J(X)$尽量大,且 X 中特征数尽量少.

由此可看出,评价函数 $J(X)$是特征选择的重要因素,它可以表示成多种形式,如分类精确度、条件概率分布或信息熵等,它的选取将直接影响选择算法的最终性能.本文研究的目的是从大量的软件特征数据集合中,选择出可作为软件胎记的特征,使选择 $J(X)$尽可能大,且 X 中特征数尽可能的少的特征集合.

2.3　聚类分析与距离度量

聚类是按照事物的某些属性,把事物聚集成簇,使簇内的对象之间具有较高的相似性,而不同簇的对象之间的相似程度较差.但是这里研究的是数据聚类,即将数据特征分为若干组或类,使得相同组内的样本距离相近,而不同组的样本距离较远.这点与特征选择的思想是一致的,因为特征选择的目标是选择一个特征子集,使得子集内部的冗余性尽量少,并与分类类别尽量相关.聚类分析算法大致可分为五大类:层次方法、划分方法、密度方法、模型方法和网格方法,其中层次聚类分析因以简单、直观明了的方式组织样本数据而成为聚类方法中最重要的一种方法[11].

距离的度量标准在聚类分析中起着关键的作用,常用的度量标准有欧氏、马氏、统计相关系数、Cosine 系数等[12].在众多度量标准中,基于信息熵或互信息的信息度量标准得到广泛研究,主要由于它能有效地表达特征间的非线性相关性,并且能以数值的形式准确量化特征的不确定性程度.许多研究实验已经表明信息度量标准在大多数情况下都能取得良好性能.由于数据挖掘所面对的数据对象日趋复杂,聚类方法的研究也面临更多新的内容和挑战,基于约束的聚类、寻找更适合的距离划分是一项具有挑战性的任务,也是一个重要的研究课题.

本文的研究也是应用互信息测量“类内”和“类间”距离.对于胎记特征的选择约束见图 1,要求同类的等价语义软件中相似度要越高越好,与其信息度量构成增益函数,与参考程序(不同类)中相似度要越低越好,与其信息度量构成惩罚函数.同时,在已选特征中,特征的冗余度要小,以保证所选择的特征既具有抗攻击性又具有唯一性.

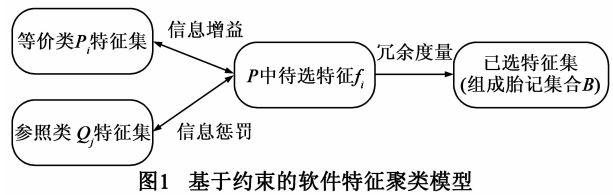

图1　基于约束的软件特征聚类模型

3 基于聚类分析的软件胎记特征选择

基于层次聚类不断将分割后的软件特征与单个候选特征进行融合,直至融合的特征数目达到指定的阈值时结束.与传统聚类算法不同的是数据点是特征,而不是样本.另外定义了两个信息度量公式分别用于计算"类内"和"类间"距离.下面详细说明算法过程.

3.1 特征聚类分析过程

设一个特征数据集 $T=(F,P_i,Q_j)$,其中 F 是原软件 P_0 的特征集合,也是特征空间,其中的特征被划分为候选类和选择类,对所有候选特征 f 进行信息度量和聚类,与已选特征类 B 计算类内距离 L_a,与等价特征集 P_i 和 Q_j 计算类间距离 L_b,其中,P_i 为原软件 P_0 经过等价语义变化后的软件(属于同类软件)的特征集,Q_j 是参考软件(属于异类软件)的特征集,在满足条件和阈值的情况下,所选择的特征集合组成胎记特征.

选择类 B 代表已选特征子集,其每个成员或数据点都对应一个已选特征 f.选择类 B 全程参与特征聚类过程,虽然聚类过程中只有一个选择类 B,但其内部成员数不止一个.候选特征 f,即目前还未被选择的特征,它在聚类过程中不断地与选择类进行组合或合并,并且每个候选类只包含一个候选特征.另外,类别 P_i 和 Q_j 也被当作参考类,它们不参与特征聚类的融合过程,而只是在聚类过程中起着"监督"作用,选择的特征距离与类别 P_i 越小越好,与类别 Q_j 越大越好,因此,通过这两类特征集合构成信息增益函数和惩罚函数.

按层次聚类分析中的合并实现方式,将特征样本根据层次结构方式进行合并,直到终止条件满足为止.首先,将由 n 个样本组成的数据集 F 分成 n 个不同的组,其中每组只包含一个样本;聚类合并时,每个候选特征 f 将分别与等价类和参考类之间计算类间距离 L_b,与已选类特征进行计算类内距离 L_a 和冗余度量.若它的 L_b 最大且 L_a 最小,那么这个候选类将与选择类进行合并,组合成新的选择类,从而得到 $n-1$ 个组.随后以相同方式继续合并,得到 $n-2$ 个组.这样一直循环下去,直到在给定阈值下,所有样本合并到同一组时算法结束.

3.2 基于互信息的距离度量

原软件 P_0 中提取的特征 f,测量其与类别之间的相关程度,即计算特征 f 与等价类 P_i 和参考类 Q_j 的类间"距离" L_b,

$$L_b(f|P_O)=\frac{1}{m}\sum_{i=m}I(f|P_o;P_i)-\beta\cdot\frac{1}{n}\sum_{j=n}I(f|P_o;Q_j) \tag{1}$$

其中 $\frac{1}{m}\sum_{i=m}I(f|P_o;P_i)$ 为 f 与同类软件距离(或信息度)均值作为增益函数,$\frac{1}{n}\sum_{j=n}I(f|P_o;Q_j)$ 为 f 与不同类软件距离(或信息度)均值作为惩罚函数.β 为调节系数,且 $\beta\in[0,1]$,调节由于数据不平衡引起的偏差.

度量候选类 f 与选择类 B 之间的距离为类内距离,与类间距离度量有所不同,因为不仅要考虑到两者间的冗余信息量,而且还要顾及信息增长率,主要原因是聚类或选择过程优先选择冗余增长率低的特征 f,那么 f 与 b 组合后,选择类(即已选子集) B 内部的冗余性程度将在一定程度内得到缓和.互信息 $I(f;b)$ 并不能表达增长率这类信息,为此这里引入关联系数的概念(又称变量相关性)作为度量冗余性的度量标准.给定两个特征 f 和 b,它们之间的相关系数为 b 在 f 已知的情况下的不确定性减少程度,即

$$\mathrm{CM}(f;b)=\frac{I(f;b)}{H(b)} \tag{2}$$

显然,关联系数 $\mathrm{CM}(f;b)$ 的取值范围是 $[0,1]$.由式(2)可知,如果 b 完全依赖于 f,那么 $\mathrm{CM}(f;b)=1$;反之,如果 b 完全独立于 f 或与 f 无关,那么 $\mathrm{CM}(f;b)=0$.关联系数 $\mathrm{CM}(f,b)$ 主要用来表示候选类 f 与单个已选特征 b 之间的"类内距离",特征聚类分析过程中候选类 f 与选择类 B 之间的"距离"也可通过对候选类 f 与选择类 B 中所有已选特征 b 的"距离"求和得到,即

$$L_a(f)=\sum_{b\in B}\mathrm{CM}(f;b) \tag{3}$$

最初 $L_a(\phi)=0$,随候选类 f 不断地与选择类 B 进行合并,类内距离 $L_a(B)$ 可按以下形式累加计算:

$$L_a(f;B)=L_a(B)+L_a(f) \tag{4}$$

除类间距离 $L_b(f)$ 和类内距离 $L_a(f)$ 外,选择类 B 的大小在特征聚类过程中也是需要考虑的因素.选择类 B 越小越好,以使后期所构造的胎记特征集不至于过于庞大,计算复杂度降低.综合以上分析,对于特征聚类分析过程的每个候选类,其评价函数是

$$J(f)=\frac{L_b(f|P_o;P_i;Q_j)}{|B|+L_a(f;B)} \tag{5}$$

其中 $|B|$ 表示选择类 B 中成员的总数,即已选特征的个数.评价标准 $J(f)$ 表示最大化特征子集和类别的相关性,同时最小化特征之间的冗余,那么 f 就具有较高的优先性与 B 进行合并,使得最终组合的选择类 B 具有较高的类区别能力.

3.3 算法实现

利用层次聚类分析,采用自底向上的合并方式,以信息度量为准则评估特征的重要程度,选择出当前性能最好的特征.具体实现是:首先对样本数据集 T 实施必要的预处理,并初始化相关参数.然后,根据互信息计算每个候选特征 f 与分类类别之间的相关程度 $I(f;$

C),若候选类 f 的 $I(f|P)$值最大,则将其加入到选择类 B 中.更新候选类 F 的类间距离 $L_b(f)$和类内距离 $L_a(f)$,以加快聚类过程的速度,这个过程一直循环迭代,直到类 B 中的特征个数超过预先设定的阈值时结束,如算法 1.

算法 1　基于聚类分析的胎记特征选择算法

Feature selection for software birthmark based on Cluster analysis

Input: A training dataset $T=(F,P_i,Q_j)$;

Output: A selected feature subset B;

Process:

1) Initialize relative parameters, e.g, $B=\emptyset, L_b=0, L_a=0$;

2) For each f in F do

Calculate its mutual information $I(f|P_o)$ with the class P_i, Q_j and $L_b(f|P_o)$;

3) $F=F-\{f\}; B=\{f\}; L_a=0$ where $f=\text{argmax}(L_b(f|P_o))$;

4) While $|B|<\delta$ do

a) For each f in F do Calculate the evaluation criterion $J(f)$ in terms of

$$J(f)=\frac{L_b(f|P_o;P_i;Q_i)}{|B|+L_a(f;B)}$$

b) Select the feature f with the maximum value of $J(f)$;

c) Combine f with B, that is, $B=B+\{f\}; F=F-\{f\}$;

d) Update the intra-distance L_b and inter-distance L_a of B, i.e., $L_a=L_a+\text{CM}(f), L_b=L_b+I(f|P_o)$;

5) return the subset B;

6) End

4　算法分析

4.1　算法效率分析

假如特征数据集 $T=(F,P_i,Q_j)$中包含 n 个样本数据和 m 个特征,那么候选类 f 与参考类之间的距离计算的时间复杂度为 $O(n)$.信息评价度量标准 $J(f)$的计算时间复杂度为 $O(nm)$.因此,每选择或组合一个候选类 f,所需要花费的时间是 $O(nm^2)$.算法总共需要循环 δ 次才终止,因而总的时间计算复杂度为 $O(\delta nm^2)$.在算法 1 中的 While 循环语句(步骤 4)将重复计算类间距离或类内距离,可以通过保留已经计算过的距离值以避免重复计算,从而达到提高算法效率的目的.

4.2　等价语义变换对聚类选择的影响

软件在受到攻击时,软件特征及其分布会发生变化,产生一系列被混淆的代码,这些代码偏离正常的聚类中心,分析等价语义变换对聚类半径及胎记特征选择的影响.

设提取的特征是一个 n 维向量 $\boldsymbol{D}(\boldsymbol{x})$,该向量是由 $D(x_1),D(x_2),\cdots,D(x_k)$共 k 个代码的特征分析得来的.由 $\emptyset(x_i)$为表征向量 $\boldsymbol{D}(\boldsymbol{x})$映射后的特征值,取 $D(x)=\frac{1}{k}\sum_{j=1}^{k}\emptyset(x_j)$,那么聚类半径 $R=\max_{1\le j\le k}\{\|\emptyset(x_i)-D(x)\|\}$,则在攻击时,产生不同的版本,要使每次的等价语义变换(仿真攻击)起到作用,在训练时使聚类半径发生偏移,使聚类中心不断发生改变.用 t 记录版本变化次数,此过程可迭代 T 次($1\le t\le T$),每次产生的变换位置个数为 n_t,那么第 t 次迭代前,特征记为$\overline{D(x_{t-1})}$,产生 n_t 个攻击点后分别记为 $Dx_1^t, Dx_2^t,\cdots,Dx_{n_t}^t$;有效攻击点 n_t',分别记为 $Dx_1^{t'}$, $Dx_2^{t'},\cdots,Dx_{n_t'}^{t'}\in\{Dx_1^t,Dx_2^t,\cdots,Dx_{n_t}^t\}$,那么 t 次后特征改变为

$$\overline{D(x_t)}=\frac{(\sum_{i=1}^{t-1}n_i)\,\overline{D(x_{t-1})}+\sum_{j=1}^{n_t}Dx_j^t}{\sum_{i=1}^{t}n_i}=\overline{D(x_{t-1})}+\frac{\sum_{j=1}^{n_t}(Dx_j^t-\overline{D(x_{t-1})})}{\sum_{i=1}^{t}n_i}$$

存在两种情况,有效语义变换和无效语义变换:(1)对有效变换满足,$Dx_j^t-\overline{D(x_{t-1})}\ge R$,($1\le j\le n_t$),每次攻击变化对特征产生改变;(2)对无效变换满足条件:$D(x)\in\{Dx_1^t,Dx_2^t,\cdots,Dx_{n_t}^t\}$, $D(x)\notin\{Dx_1^{t'},Dx_2^{t'},\cdots,Dx_{n_t'}^{t'}\}$且 $Dx_j^t-\overline{D(x_{t-1})}<R$.

由此可见,当 $n_t=n_{t'}$时,等价语义变换(仿真攻击)效果最好,可记为 $Dx_j^t=\overline{D(x_{t-1})}+R$,以下取 R 为初始半径,值为 0.

定理 1　在等价语义变换(仿真攻击)中,记 $M_{t-1}=\sum_{i=1}^{t-1}\sum_{j=1}^{n_i}Dx_j^i, N_{t-1}=\sum_{i=1}^{t-1}n_i, m_t=\sum_{i=1}^{n_t}Dx_j^t$,若满足$\frac{m_t}{M_{t-1}}-\frac{n_t}{N_{t-1}}>0$,则有$\overline{D(x_t)}>\overline{D(x_{t-1})}$,即攻击后半径不断增大.

证明　攻击后半径不断增大,即存在$\overline{D(x_t)}-\overline{D(x_{t-1})}>0$ 成立,由于 $\overline{D(x_t)}=\frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}Dx_j^i}{\sum_{i=1}^{t}n_i}$, $\overline{D(x_{t-1})}=\frac{\sum_{i=1}^{t-1}\sum_{j=1}^{n_i}Dx_j^i}{\sum_{i=1}^{t-1}n_i}$,代入上式可得:$\overline{D(x_t)}-\overline{D(x_{t-1})}=\frac{\sum_{i=1}^{t}\sum_{j=1}^{n_i}Dx_j^i}{\sum_{i=1}^{t}n_i}-\frac{\sum_{i=1}^{t-1}\sum_{j=1}^{n_i}Dx_j^i}{\sum_{i=1}^{t-1}n_i}=\frac{N_{t-1}(M_{t-1}+m_t)-M_{t-1}(N_{t-1}+n_t)}{N_{t-1}(N_{t-1}+n_t)}>0$,所以可得$\frac{m_t}{M_{t-1}}-\frac{n_t}{N_{t-1}}>0$满足.因此,只要按照一定的语义变换策略和仿真攻击,能够使聚类中心不断偏移,半径不断增大,从而达到识别软件多态变化的目的.

5 总结

本文提出的是一种过滤式(filter)特征选择算法,适用于不同形式(字符串、向量、集合、树型和图型)的软件胎记选择,实用性较强.该特征选择算法,可与其它选择方法结合,如程序切片和聚类相结合,过滤式(Filter)和封装式(Wrapper)相结合,以达到更好的胎记特征选择效果.

参考文献

[1] Mahmood Y, Sarwar S, Pervez Z, et al. Method based static software birthmarks: A new approach to derogate software piracy[A]. Proceedings of the 2nd International Conference on Computer, Control and Communication[C]. Karachi: IEEE, 2009.149 – 155.

[2] Xin Z, Chen H, Wang X, et al. Replacement attacks on behavior based software birthmark[J]. Information Security, 2011, 11(5): 1 – 16.

[3] Myles G, Collberg C. K-gram based software birthmarks[A]. Proceedings of the symposium on Applied computing[C]. New York: ACM, 2005.314 – 318.

[4] Lu B, Liu F, Ge X, et al. Feature n-gram set based software zero-watermarking[A]. Proceedings of the International Symposiums on Information Processing[C]. Moscow: IEEE, 2008.607 – 611.

[5] Danicic S, Hierons R M, Laurence M R. On the computational complexity of dynamic slicing problems for program schemas [J]. Mathematical Structures in Computer Science, 2011, 21(06): 1339 – 1362.

[6] 刘巍伟,石勇,郭煜,等.一种基于综合行为特征的恶意代码识别方法[J].电子学报,2009,37(4):696 – 700.
Liu Wei-wei, Shi Yong, Guo Yu, et al. A malicious code detection method based on integrated bethavior characterization[J]. Acta Electronica Sinica, 2009, 37(4): 696 – 700. (in Chinese)

[7] Tamada H, Okamoto K, Nakamura M, et al. Dynamic software birthmarks to detect the theft of windows applications[A]. Proceedings of the 8th International Symposium on Future Software Technology[C]. Xi'an, China: Tokyo Software Engineers Association, 2004.37 – 43.

[8] Proietti M, Pettorossi A. Semantics preserving transformation rules for Prolog[A]. Proceedings of the SIGPLAN symposium on Partial evaluation and semantics-based program manipulation[C]. New York: ACM, 1991.274 – 284.

[9] Tamada H, Nakamura M, Monden A, et al. Design and evaluation of birthmarks for detecting theft of java programs[A]. Proceedings of the International Conference on Software Engineering[C]. Austria: ACTA, 2004.569 – 575.

[10] 鲍明,管鲁阳,李晓东,等.基于欧氏距离分布熵的特征优化研究[J].电子学报,2007,35(3):469 – 473.
Bao Ming, Guan Lu – yang, Li Xiao-dong, et al. A study on optimum classification character based on the distributive entropy of euclidian distance[J]. Acta Electronica Sinica, 2007, 35(3): 469 – 473. (in Chinese)

[11] Gruvaeus G., Wainer H. Two additions to hierarchical cluster analysis[J]. British Journal of Mathematical and Statistical Psychology, 2011, 25(2): 200 – 206.

[12] 蒋盛益,郑琪,张倩生.基于聚类的特征选择方法[J].电子学报,2008,36(12A):157 – 162.
Jiang Sheng-yi, Zheng Qi, Zhang Qian-sheng. Ckyster-based feature selection[J]. Acta Electronica Sinica, 2008, 36(12A): 157 – 162. (in Chinese)

作者简介

罗养霞 女,1974 年 1 月出生,陕西户县人.西安财经学院信息学院任教,讲师.西北大学信息学院,在读博士生,从事软件保护、水印、胎记方面的研究.
E-mail: yxluo8836@163.com

房鼎益 男,1959 年 3 月出生,陕西汉中人.西北大学信息学院教授、博士生导师.主要研究方向为网络与信息安全、无线传感器网络及其应用.