ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2007/01(03)-0282-11

E-mail:fcst@public2.bta.net.cn
http://www.ceaj.org
Tel:+86-10-51616056

基于树编辑距离的层次聚类算法*

乔少杰[1,2],唐常杰[1+],陈 瑜[1],彭 京[3],温粉莲[1]
QIAO Shaojie[1,2], TANG Changjie[1+], CHEN Yu[1], PENG Jing[3], WEN Fenlian[1]

1.四川大学 计算机学院,成都 610065
2.新加坡国立大学 计算机学院,新加坡 117590
3.北京大学 信息科学技术学院,北京 100871
1.School of Computer, Sichuan University, Chengdu 610065, China
2.School of Computing, National University of Singapore, 117590, Singapore
3.School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
+Corresponding author: E-mail: tangchangjie@cs.scu.edu.cn

QIAO Shaojie, TANG Changjie, CHEN Yu, et al. A new hierarchical clustering algorithm based on tree edit distance. Journal of Frontiers of Computer Science and Technology, 2007,1(3):282-292.

Abstract: In order to recognize the false status which has been forged and tempered by suspects, a new method is proposed to compute attribute similarities based on tree edit distance, and its mathematical properties are proved. The paper proposes a new clustering algorithm based on hierarchical encoding method named HCTED(Hierarchical Clustering Algorithm Based on Tree Edit Distance). This method uses tree edit distance to compute attribute similarities with minimum cost, overcomes the shortage of traditional clustering algorithms and improves the precision of clustering according to the predefined threshold. Experiments demonstrate that the new method is accurate and efficient in identity recognition, discuss the effects of different experimental parameters, and show that HCTED is more accurate and faster than traditional clustering algorithms. The new algorithm has been used in data analysis of transient population for public security successfully.

Key words: tree edit distance; hierarchical clustering; attribute similarity; data mining

摘 要:为了识别犯罪嫌疑人伪造和篡改的虚假身份,利用树编辑距离计算个体属性相似性,证明了树编辑距离的相关数学性质,对属性应用层次编码方法,提出了一种新的基于树编辑距离的层次聚类算法 HCTED(Hi-

* the National Natural Science Foundation of China under Grant No.60773169,60473071(国家自然科学基金);the Postdoctoral Science Foundation of China under Grant No.20060400002(中国博士后科学基金);the Youth Science and Technology Foundation of Sichuan Province of China under Grant No.2007Q14-055,08JJ0109(四川省青年科技基金).
Received 2007-08,Accepted 2007-10.

erarchical Clustering Algorithm Based on Tree Edit Distance)。新算法通过树编辑操作使用最少的代价计算属性相似性,克服了传统聚类算法标称型计算的缺陷,提高了聚类精度,通过设定阈值对给定样本聚类。实验证明了新方法在身份识别上的准确性和有效性,讨论了不同参数对实验结果的影响,对比传统聚类算法,HCTED 算法性能明显提高。新算法已经应用到警用流动人口分析中,取得了良好效果。

关键词:树编辑距离;层次聚类;属性相似性;数据挖掘

文献标识码:A　　**中图分类号**:TP311

1　引言

在犯罪活动中，犯罪嫌疑人经常通过隐瞒或伪造自己真实身份的方式进行欺诈等犯罪活动，因此身份识别在社会安全管理中有着极其重要的作用。将数据挖掘技术应用于犯罪和恐怖组织成员身份识别成为相关研究的热点[1,2]。犯罪嫌疑人或恐怖分子利用假身份进行的违法行为和恐怖活动常常对社会造成极大的损失，因此如何进行准确有效的身份识别成为安全部门和情报机关亟待解决的问题。借助计算机进行身份识别是该领域的新兴技术，这项技术可广泛应用于反恐以及犯罪过程中可疑人物的身份鉴定[3]。

传统通过手工进行身份鉴别会耗费巨大的人力和物力,此外,许多主观不确定因素对准确有效识别嫌疑人身份具有很大的影响,如个人喜好、情感、经验等因素。使用计算机搜索历史记录有助于准确识别身份，但使用完全精确匹配很难识别出真正的犯罪嫌疑人。

文献[4]表明,为了避免暴露真实身份,犯罪嫌疑人往往伪造部分虚假的身份信息，和历史记录比较相似,但不完全一致,导致采用传统精确匹配方法无法准确识别其身份。因此,文章提出了基于树编辑距离的层次聚类算法用于身份识别，算法的基本思想是通过计算给定人员档案信息与历史数据的相似度来鉴别用户身份,如果相似度大于某个阈值,则认为和历史记录相匹配,否则不匹配。通过这种方法可以有效地从人事档案中识别出虚假信息并予以排除。由于文章利用编辑距离对人员身份进行聚簇处理,即使犯罪分子人为捏造出部分混淆信息，经过聚类方法也可以有效地对其鉴定识别。

2　树编辑距离

2.1　基本概念

文章引入树编辑距离计算属性值间的相似度,为了便于理解,首先给出编辑距离的定义及其性质[5,6]。

定义 1　编辑距离(Edit Distance) 记为 $d(S_1,S_2)$,其中 S_1,S_2 为字符串。编辑距离被定义为使 S_1 和 S_2 成为相同字符串需要以下操作的最小次数。

操作 1　把某个字符 ch_1 变成 ch_2;

操作 2　删除某个字符;

操作 3　插入某个字符。

例如:对于字符串 S_1="12433"和 S_2="1233",则可以通过在 S_2 中间插入 ch=4 得到"12433",和 S_1 一致,即 $d(S_1,S_2)$=1(进行了一次插入操作)。

性质 1　计算两个字符串 S_1+ch_1 和 S_2+ch_2 的编辑距离存在如下性质(其中|S|表示字符串 S 中的字符长度):

(1)$d(S_1,\text{“ ”})=d(\text{“ ”},S_1)=|S_1|$;

(2)$d(\text{“}ch_1\text{”},\text{“}ch_2\text{”})=ch_1==ch_2?0:1$;

(3)$d(S_1+ch_1,S_2+ch_2)=\min(d(S_1,S_2)+ch_1==ch_2\text{” }?\ 0:1,d(S_1+ch_1,S_2),d(S_1,S_2+ch_2))$。

性质 1 表明:由于定义了 3 种操作来作为编辑距离的一种衡量方法，那么对 ch_1,ch_2 的可能操作仅有:

(1)把 ch_1 变成 ch_2;

(2)S_1+ch_1 后删除 ch_1,即 $d=(1+d(S_1,S_2+ch_2))$;

(3)S_1+ch_1 后插入 ch_2,即 $d=(1+d(S_1+ch_1,S_2))$。

操作 2 和操作 3 又可以等价为:

(1)S_2+ch_2 后添加 ch_1,即 $d=(1+d(S_1,S_2+ch_2))$;

(2)S_2+ch_2 后删除 ch_2,即 $d=(1+d(S_1+ch_1,S_2))$。

由于每个字符串均可通过插入一个根结点构成一棵深度为 2 的树，因此可以将计算编辑距离问题转化为计算由两个字符串构成的树之间相似度的问题,定义 2 给出树编辑距离的形式化描述[7]。

定义 2 树编辑距离(Tree Edit Distance)记为 $d(T,T')$,其中 T 和 T'表示树,$|T|$表示树中结点个数,$T[i]$表示通过遍历得到位置为 i 的结点,树编辑距离定义为 $d(T,T')=\min\{r(P)|P$ 是将 T 转化为 T'的一系列树编辑操作$\}$,树编辑操作包括:

(1)插入操作:插入一个结点到一棵树中;

(2)删除操作:从一棵树中删除一个结点;

(3)转换操作:将一棵树中的结点转化为其他结点。

2.2 树编辑距离求解

计算树编辑距离的过程是求使 T 转化为 T'所需树编辑操作的最少次数，文中将这一求解过程定义为映射[7]。如图 1 所示,从 $T[i]$到 $T'[j]$的连线表示:

(1)如果 $T[i]\neq T'[j]$,则 $T[i]$转化为 $T'[j]$;

(2)如果 $T[i]=T'[j]$,则 $T[i]$不变化。

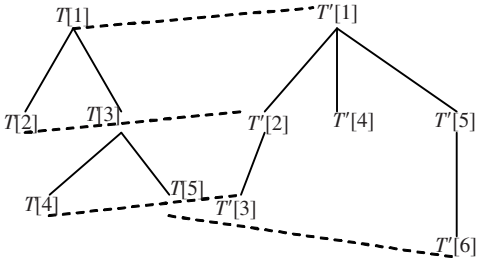

图 1 树映射关系图

Fig.1 Tree mapping diagram

T 和 T'中没有用线连接的结点分别表示要删除的结点和插入的结点,文中将整个过程称为映射,其形式化定义如下。

定义 3 映射(mapping)记为一个三元组(M,T,T'),其中 M 是满足如下关系的整数对(i,j):

(1)$1\leqslant i\leqslant|T|,1\leqslant j\leqslant|T'|$;

(2)对于 M 中的两对数(i_1,j_1)和(i_2,j_2):

①$i_1=i_2$ 当且仅当 $j_1=j_2$;

②$i_1<i_2$ 当且仅当 $j_1<j_2$;

③$T[i_1]$是 $T[i_2]$的祖先或后继结点当且仅当 $T'[j_1]$是 $T'[j_2]$的祖先或后继结点。

令 M 为从 T 到 T'的一个映射,I 和 J 分别为 T 和 T'中未连接的结点,将映射代价[7]定义为:

$$cost(M)=\sum_{(i,j)\in M}(T[i]-T'[j])+\sum_{(i\in I)}r(T[i]\rightarrow\Lambda)+\sum_{(j\in J)}r(\Lambda\rightarrow T[j]) \quad (1)$$

为了证明 $d(T,T')$可以通过最小的映射代价得到,引入如下两个引理[7]。

引理 1 令 M_1 为 T_1 到 T_2 的映射,M_2 为 T_2 到 T_3 的映射,则:

(1)$M_1\circ M_2=\{i,k|\exists j,(i,j)\in M_1,(j,k)\in M_2\}$表示从 T_1 到 T_3 的映射;

(2)$cost(M_1\circ M_2)\leqslant cost(M_1)+cost(M_2)$。

证明 令(i_1,k_1)和(i_2,k_2)为 $M_1\circ M_2$ 中的两条连线,存在 j_1 和 j_2 使得$(i_1,j_1),(i_2,j_2)\in M_1,(j_1,k_1),(j_2,k_2)\in M_2$,根据映射的定义有:

①$i_1=i_2$ 当且仅当 $j_1=j_2,j_1=j_2$ 当且仅当 $k_1=k_2$;

②$i_1<i_2$ 当且仅当 $j_1<j_2,j_1<j_2$ 当且仅当 $k_1<k_2$;

③$T_1[i_1]$是 $T_1[i_2]$的祖先或后继结点当且仅当 $T_2[j_1]$是 $T_2[j_2]$的祖先或后继结点,$T_2[j_1]$是 $T_2[j_2]$的祖先或后继结点当且仅当 $T_3[k_1]$是 $T_3[k_2]$的祖先或后继结点。因此 $M_1\circ M_2$ 是从 T_1 到 T_3 的映射关系。

引理 1 的证明参见文献[7],这里略去。

引理 2 对于从 T 转化到 T'的一系列的树编辑距离操作 S，存在从 T 到 T'的映射关系 M，使得 $cost(M)\leqslant r(S)$。

证明 如果 $S=s_0,s_1,\cdots,s_m$ 是一系列树编辑操作,$T_0,T_1,\cdots,T_m$ 是从 T_0 通过 S 操作转化到 T_m 的一系列的树,根据文献[8]中给出的定理 1 可以知道存在从 T_0 到 T_m 的映射 M,满足 $cost(M)\leqslant r(S)$。

根据上述引理可以得到树编辑距离计算的最小代价定理。

定理 1 最小代价定理 $d(T,T')=\min\{cost(M)|M$ 是从 T 到 T'的映射$\}$

证明略,参见文献[7]中定理 1 的证明。

文中对传统的编辑距离进行改进，提出了一种基于树编辑距离的聚类算法，该算法的优势在于能够通过最少的代价计算字符串之间的距离。第 4 章将给出利用树编辑距离计算不同属性相似度的方法,进而对个体进行聚类。

3 聚类分析

聚类分析目的是寻找数据中潜在的自然分组结构和感兴趣的关系。对样本空间进行聚类,须规定样本相异度值,其由相异度矩阵[9]表达,定义如下。

定义 4 相异度矩阵(dissimilarity matrix) 用来存储 n 个样本两两之间的相似性,形式为 $n\times n$ 维矩阵:

$$\begin{bmatrix} 0 & \cdots & & \\ d(X_2,X_1) & 0 & \cdots & \\ d(X_3,X_1) & d(X_3,X_2) & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ d(X_n,X_1) & d(X_n,X_2) & \cdots & 0 \end{bmatrix}$$

其中 $d(X_i,X_j)$是样本 X_i 和 X_j 相异性量化描述,取值为非负数。

解释:对象 i 和 j 越相似,其值越接近 0;否则其值越大。显然,相异度值是样本之间距离的度量参考。

给定一数据样本集 $X=\{X_1,X_2,\cdots,X_n\}$,其中 $X_i=\{X_{i1},\cdots,X_{ik}\}$和 $X_j=\{X_{j1},\cdots,X_{jk}\}$是两个拥有 k 个属性的样本,每个属性用第 4 章介绍的层次编码方法表示,$d(X_i,X_j)$表示第 i 个样本与第 j 个样本间的距离。聚类算法先定义一个适当度量尺度，据此计算任意两个样本之间的距离。当两个样本之间的距离小于某个阈值 d_0 时,将两个样本划归为同一类。阈值 d_0 影响簇的数量和规模:d_0 越小,簇规模越小,簇数目越多,反之类似。如果 d_0 太大,所有样本将被分为同一簇;如果 d_0 太小,每个样本将单成一簇。因此选取适当的 d_0 值是进行成功聚类的关键，将在第 6 章对其取值进行讨论。

设 C_i 和 C_j 分别代表两个不同的簇,X_i,X_j 分别表示 C_i 和 C_j 的样本代表(簇核心),则簇 C_i 和 C_j 之间距离度量标准如下[10]:

(1)最小距离:$D_{\min}(C_i,C_j)=\min|X_i-X_j|$,其中 $X_i\in C_i,X_j\in C_j$;

(2)最大距离:$D_{\max}(C_i,C_j)=\max|X_i-X_j|$;

(3)中间距离:$D_{\mathrm{mid}}(C_i,C_j)=|m_i-m_j|$,其中 m_i 和 m_j 是 C_i 和 C_j 的质心,即中间样本代表;

(4)平均距离:

$$D_{avg}(C_i,C_j)=\frac{1}{n_in_j}\sum\sum|X_i-X_j| \tag{2}$$

其中,n_i 和 n_j 表示类 C_i 和 C_j 样本数目。

4 属性距离计算

犯罪嫌疑人经常通过谎报姓名、出身年月、身份证号、职业以及地址等手段隐瞒真实身份。可能手法包括:对于名字将真实的姓加虚假的名字,虚假的姓加真实的名字,或者结合二者;对出身年月采用增减年月份;对身份证号变换其中几位数字;对地址是用假的大地址和真实街道的组合等。

标识罪犯的属性,如出生年月、身份证号码、行政区划、地址等都是层次编码变量,其特点是可以分为 n 个互不相交的部分,记为 $T=(a_1,a_2,\cdots,a_n)$,其中每个属性值是一个标称型变量，同时任意 a_i 都有一个父结点 a_{i-1},a_i 的值由其父结点决定。

文章采用文献[11]提出的层次编码平衡树对每条记录编码，用叶子来代表任意层次编码变量 P,记为 $T(P)$,平衡树 T 上每层结点之间的路径权重记为 $\omega_1,\omega_2,\cdots,\omega_n(\omega_i>0,i=1,2,\cdots,n-1)$,其中 ω_i 表示第 i 到 $i+1$ 层结点间的权重，取值代表结点之间的相异程度,越小表示越相似。以行政区划为例,成都市武侯区的行政区划(510107)可以用图 2 的粗线路径表示[11]。

从层次编码变量的特点上看,它具有树的特点,可以通过遍历树得到不同记录的编码。文中应用树编辑距离求解两条记录中属性间的距离，从而确定属性间的相似度。为了方便计算,对每个属性距离进

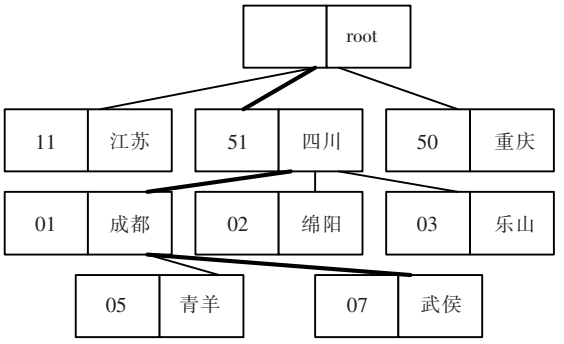

图 2 层次编码平衡树

Fig.2 Hierarchical balance tree

行归一化处理,如下式所示:

$$d_x=\frac{D_x}{\max(|S_1|,|S_2|)} \tag{3}$$

其中,D_x 表示两条记录在属性 x 上的编辑距离,d_x 表示归一化后的编辑距离,分母表示两个属性值的最长串长度,d_x 为一个介于 0 和 1 之间的实数值。

在各属性距离分量基础上使用加权欧氏距离,如下式所示:

$$d(i,j)=\sqrt{(w_1\cdot d(x_{i_1},x_{j_1}))^2+\cdots+(w_n\cdot d(x_{i_n},x_{j_n}))^2} \tag{4}$$

其中,$d(i,j)$表示两条记录间的编辑距离,$d\in[0,1]$;$w_1,w_2,\cdots,w_n$ 分别表示每个属性维度上的权重,如果 $w_1=w_2=\cdots=w_n=1$,则式(4)退化为传统的欧式距离。

5 基于树编辑距离的层次聚类算法

文章提出的基于树编辑距离的层次聚类算法 HCTED(Hierarchical Clustering Algorithm Based on Tree Edit Distance)是一种改进的层次聚类算法,该算法将树编辑距离作为个体相似性度量标准,将层次编码方法应用到聚类算法中,根据给定的阈值对给定身份标识划归分类,从而得到以相似性为基本度量的簇。由于簇的数目无法事先确定,需要通过设置阈值对样本进行聚类。

HCTED 算法要点为:设 w 为窗口宽度值,表示进行比较的临近记录范围;q 表示一个队列,队列大小设置为 w,用来存储聚簇的代表值,最多存放 w 个。当簇的个数大于 w 时,第一个元素出队,直到所有的记录聚类完成。如果队列 q 不为空,则计算当前记录 r_i 与 q 的分类代表值的距离:

(1)若距离小于阈值,则将 r_i 并入 C_j;

(2)否则,计算与已经分类的不在 q 队列中记录的距离,如果距离小于阈值,则并入该记录属于的分类中,否则创建新的簇。

基于树编辑距离的层次聚类算法实现如下。

算法 1 基于树编辑距离的层次聚类算法

输入:窗口宽度 w,阈值 *threshold*;

输出:聚类簇的个数 *cluster_num*。

```
(1)Set w; //设置窗口宽度 w 值
(2)Encode each record by calling function HEncode(); //对每条记录进行层次编码
(3)for each record
//如果不属于某个类,调用函数 Tedist()计算记录 r_i 与现有聚簇代表值的距离,如果和某个簇代表值的距离小于阈值,则加入该簇;否则创建新簇。
(4)  if r_i not in C_j
(5)    dist[i]=Tedist(r_i,C_j,len(r_i),len(C_j));
(6)    if dist[i]≤threshold then
(7)      union(r_i,C_j);
(8)    else //如果队满,则第一个分类出队
(9)      if sizeof(q)=w then
(10)       q.dequeue();
(11)     end if
(12)     Create a new cluster C_k;
(13)     cluster_num++;
(14)     q.queue(C_k);
(15)     union(r_i,C_k);
(16)   end if
(17)  end if
(18) end for
(19) return cluster_num;
```

定理 2 设记录数为 n,树的长度分别为 p 和 q,l 和 l' 表示两棵树的最大深度,则算法 1 时间复杂度为 $O(n^2*p*q*l^2*l'^2)$。

证明 利用算法 1 对每条记录进行聚类,如果不在已有簇中,则需要计算当前记录和已有各个簇间的距离,对 n 条记录需要计算 $\frac{n(n+1)}{2}$ 次距离。根据文献[7]中定理 5.1 可知,每次计算属性距离所需时间为 $p*q*l^2*l'^2$,因此算法 1 的时间复杂度为 $O(n^2*p*q*l^2*l'^2)$。

阈值 *threshold* 的设置会影响聚类效果,因此在进行每次聚类计算前需对其进行设置,这将在下一章实验部分详细说明。

6 实验

6.1 实验环境及数据集

文中提出的聚类算法 HCTED 在 Sybase PowerBuilder 9.0 平台上实现,数据库为 Microsoft SQL Server 2000,操作系统为 Windows XP Professional。硬件环境为:Intel P4 CPU 2.4 GHz,512 MB 内存。

实验数据集来源于某地近两年收集的流动人口旅店住宿登记信息,总数据量为 3 700 000 条,数据集大小约为 4 GB。每条记录涉及到 21 个属性,其中包括:姓名,性别,国籍,出生日期,身份证号,行政区划,入住时间,房间号,退房时间等。

由于并非所有属性均对犯罪嫌疑人身份鉴别起决定作用,故应找出对身份鉴定起决定作用的关键属性,然后对这些属性进行综合比较进而分析出两条记录的相似性。通过反复实验以及与犯罪心里学专家讨论,作者发现名字、出身日期、身份证号和地址这 4 个属性对辨识身份起关键作用,因此将其作为评价相似度的属性集合,相对而言其他属性影响较小,故可以忽略。实验聚类簇的数目设置为 10,实验前需将指定数据全部读入内存,并完成相应的预处理工作,文中所述实验得到的运行时间不包括数据加载和预处理的时间。

6.2 实验描述

实验的目的是验证基于树编辑距离的层次聚类算法 HCTED 在有效性和效率等方面的性能优势,实验数据来源于某地相关部门提供的原始数据集。HCTED 算法已经被成功地应用到犯罪数据挖掘系统 CrimeMiner 中,该系统用于辅助流动人口信息分析,系统界面如图 3 所示,其中包括:数据预处理,数据挖掘(该模块集成了 HCTED 算法,分类,关联规则分析),社会网络分析(包括网络信息初始,网络可视化,关键人物查找功能)和数据信息查询模块。用户通过设置数据更新时间,聚簇的数目和阈值可以得到如图 4 所示的运行结果。图 4 中类标用于标识不同聚簇,具有相同类标的数据记录归为一类。

图 3 身份识别功能界面

Fig.3 Interface of identity recognition

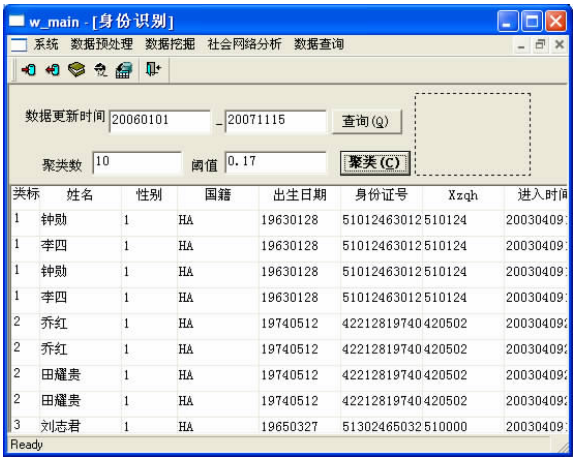

图 4 HCTED 算法聚类结果图

Fig.4 Results of clustering by HCTED

文中采用通用的聚类评价度量标准来验证 HCTED 算法的准确性,即查准率(precision)和查全

率(recall)[10]。

设 D 为实体集合,该集合中包含 m 个唯一个体且每个个体至少对应一个实体,$d_{ij}(1\leqslant i\leqslant m,j\geqslant 1)$代表第 i 个集合的第 j 个实体,算法将所有的实体分为 n 个簇,每个簇记作 $C_i=\{d_{ij}\in D,d_{ij}$ 关联 k 个个体$\}$,其中 $k=1,2,\cdots,n$,则每个簇具有如下属性:

$$C_i\cap C_j=\phi \text{ 且 } \cup C_i=D \tag{5}$$

例如簇 C 对应 3 个不同实体$\{c_1,c_2,c_3\}$,簇 F 对应 2 个不同实体$\{f_1,f_2\}$。假设利用聚类算法得到 2 个簇$\{c_1,c_2\}$和$\{c_3,f_1,f_2\}$,则根据表 1 所示的混淆矩阵(Confusion Matrix)[12]得到公式(6)和(7):

$$precision=\frac{TP}{TP+FP} \tag{6}$$

$$recall=\frac{TP}{TP+FN} \tag{7}$$

在本例中,查准率 $precision=2/4=50\%$,查全率 $recall=2/4=50\%$。

表 1 混淆矩阵

Table 1 Confusion matrix

	Predicted Cases	Negative Cases
Predicted Positive	True Positive(TP)	False Positive(FP)
Predicted Negative	False Negative(FN)	True Negative(TN)

6.3 HCTED 算法有效性分析

实验 1 不含属性空缺值的个体识别

实验目的是证明 HCTED 算法的准确性,实验数据不含属性空缺值,取阈值 *threshold* 为 0.19,实验结果如表 2 和表 3 所示。

表 2 不含空缺值的个体识别正确率

Table 2 Accuracy of identity recognition without missing values

记录数	个体数	识别个体数	正确率
30	15	15	100%
70	36	36	100%
100	51	49	96.1%
500	121	118	97.5%
1 000	218	212	97.2%
1 500	543	525	96.7%

表 3 查准率与查全率

Table 3 Result of precision and recall

记录数	查准率	查全率
30	100%	100%
70	100%	100%
200	98.2%	90.1%
700	90.2%	80.2%
1 500	86.6%	72.3%
平均	95.0%	88.5%

表 2 表明在不含属性空缺值的情况下,算法识别正确率较高;从表 3 可以发现 HCTED 算法的查准率和查全率均较高,从而验证了 HCTED 算法的有效性。

实验 2 包含属性空缺值的个体识别

实验目的是分析不同比例的属性空缺值对个体识别率的影响,选取的样本包含不同比例的属性空缺值,取 *threshold* 为 0.18,样本数据包含属性空缺值比例如表 4 所示,实验结果如表 5 所示。

表 4 属性空缺值比例

Table 4 Ratio of missing values

空缺属性	记录数	百分比
无空缺属性	1 000	81.30%
姓名	10	0.81%
地址	107	8.70%
身份证号	35	2.85%
出生日期	78	6.34%
总数	1 230	100%

表 5 含空缺值的个体识别正确率

Table 5 Accuracy of identity recognition with missing values

记录数	个体数	识别个体数	正确率
1 230	302	263	87.1%

含有空缺值属性的情况恰好与实际吻合,因为真实的样本数据往往都包含不同程度的噪声数据。表 5 表明包含比例不同的属性空缺值记录的个体识别率有所下降,原因在于如果存在属性空缺值,则计算出不同个体之间的编辑距离较大,进而产生实验误差,但总体识别率仍然很高,说明 HCTED 算法具有一定的抗噪声能力。

实验 3 不同属性空缺对正确率影响

实验目的是分析不同属性空缺值对个体识别正确率的影响。由于姓名、出生日期、身份证号、地址这 4 个属性对计算 2 条记录之间的相似度影响较大,因此选取这 4 个属性包含空缺值的记录,得到属性空缺值对识别率的影响,如图 5 所示。

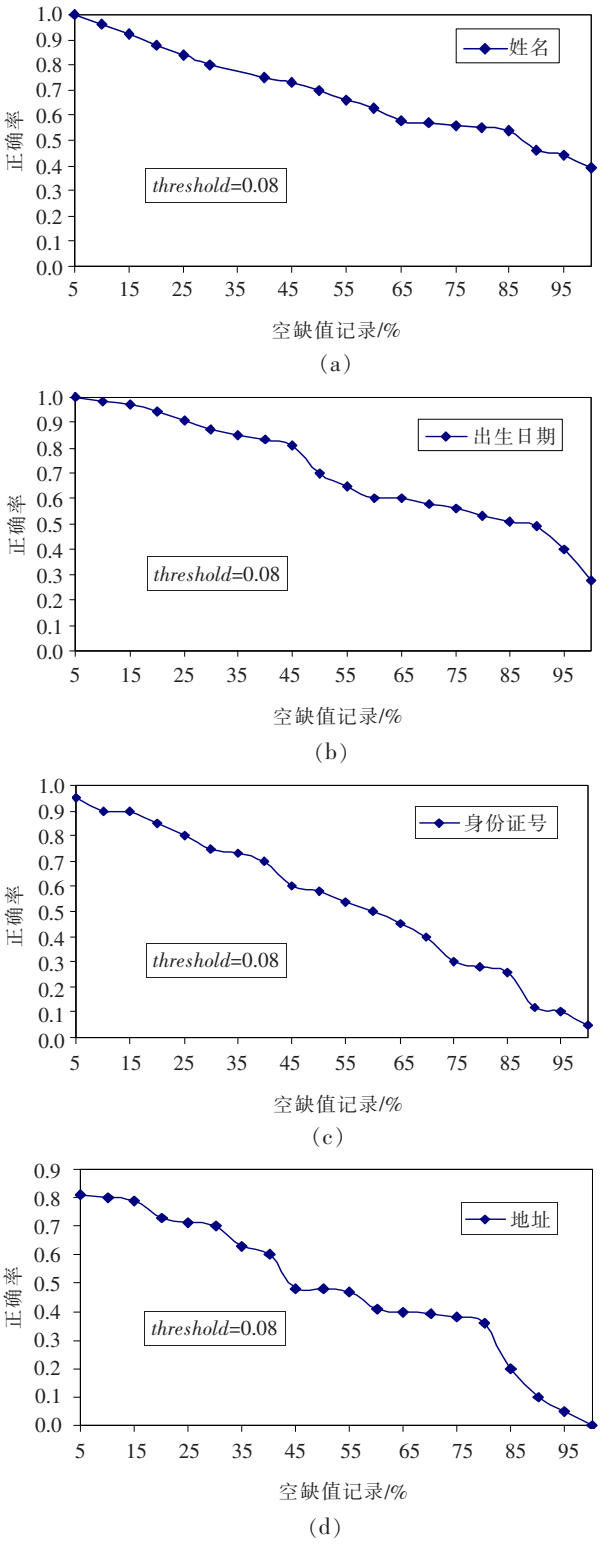

图 5 属性空缺对个体识别率的影响

Fig.5 Effect on accuracy by missing values

通过实验可以得到结论:(1)随着属性空缺值增大,识别准确率逐渐下降,这恰好与实际相符;(2)地址和身份证号在计算样本间距离时起重要作用,这两个属性的缺失在很大程度上影响个体识别率。

实验 4 阈值与识别正确率的关系

实验目的是研究阈值与个体识别率之间的关系,取 1 000 条记录,其中包含 754 个唯一个体,实验结果如图 6 所示。

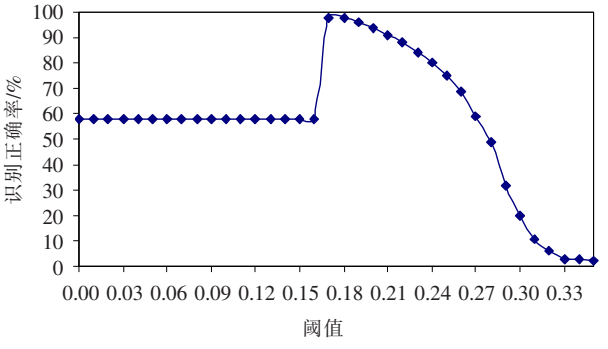

图 6 阈值与个体识别正确率的关系

Fig.6 Relationship between threshold and recognition accuracy

当样本数目改变时,实验结果与图 6 基本一致,因此这里仅给出样本数目为 1 000 的实验结果。图 6 表明在阈值达到 0.17 前,个体识别率基本不变。当阈值进入某一个区间(0.17~0.18)时,识别正确率达到最大值,之后随着阈值增大,识别率随之下降。通过图 6 可以发现:在阈值达到某个最佳值(识别率最大的值)前,其大小与识别率没有关系;当超过这个值后,阈值和个体识别率之间呈正态分布。

6.4 HCTED 算法效率分析

实验 5 阈值对运行时间的影响

实验目的是分析阈值与算法运行时间的关系,选取 1 000 条记录,其中包含 754 个唯一个体,实验结果如图 7 所示。

图 7 表明,当阈值到达最佳值(此处为 0.17)前,算法运行时间波动不大。当阈值大于该值后,算法运

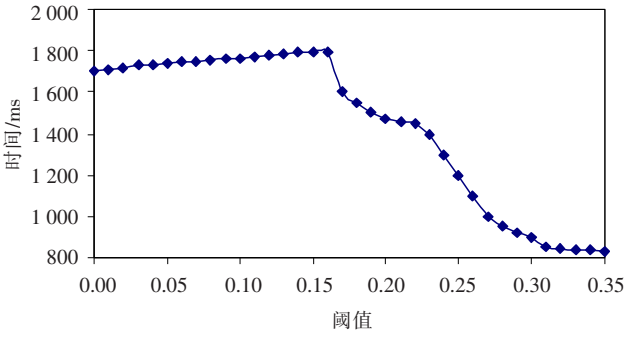

图 7 阈值与算法运行时间关系

Fig.7 Relationship between threshold and running time

行时间值随着阈值的增加反而下降,原因在于:实验中包含的样本数量一定,未到达阈值最佳值前,聚类结果相同;超过最佳阈值时,阈值越大,分类越粗糙,所需时间越少。

实验 6 样本数与阈值关系

实验目的是挖掘样本数目与最佳阈值(能使个体识别率达到最大的阈值)之间的关系,实验结果如表 6 所示。

表 6 样本数目与最佳阈值关系

Table 6 Relationship between the number of samples and the optimal threshold

样本数	唯一个体数	最佳阈值
50	50	任意值
50	40	0.15
100	75	0.18
200	180	0.08
400	230	0.23
1 000	754	0.17

表 6 表明,当样本中每条记录都是唯一个体时其最佳阈值可以取任意值,其他情况下最佳阈值的选取没有规律,需要通过多次实验确定具体取值,大量实验表明最佳阈值的选取介于 0.15~0.25 之间。

6.5 海量数据下 HCTED 算法性能比较

为了验证数据规模为海量时 HCTED 算法的性能,实验中,选取 1 000~10 000 条数据,对 HCTED 算法和朴素聚类算法的运行时间进行比较,实验结果如图 8 所示,其中横坐标表示数据记录大小,纵坐标为运行时间。

实验表明,在 4 000 条以下的数据规模 HCTED 算法消耗的时间非常小,可以忽略不计,即使数据增大到 10 000 条,运行时间与朴素算法相比依然很小;而朴素算法的时间增长很快,原因在于文中采用层次编码的方法,压缩了字符串编码长度,减少了运行时间。从图 8 可以发现 HCTED 算法性能远远高于朴素算法。

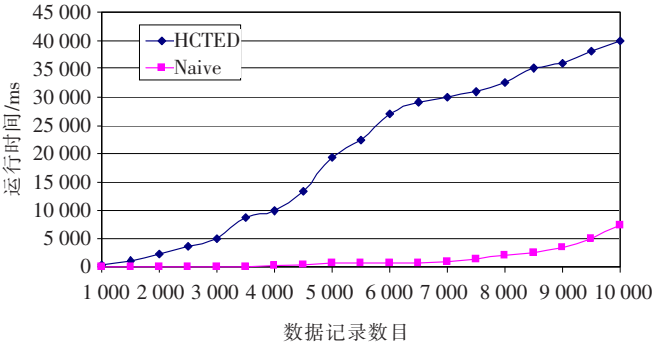

图 8 HCTED 与朴素算法运行时间对比

Fig.8 Comparison of running time between HCTED and naive clustering algorithm

为了进一步证明 HCTED 算法的聚类效果优于传统的聚类算法,实验比较 HCTED 与 k-Means 算法[10]的聚类准确性,其中阈值取值为 0.2,结果如图 9 和图 10 所示,其中横坐标表示数据记录数目/500,纵坐标分别表示查准率和查全率。

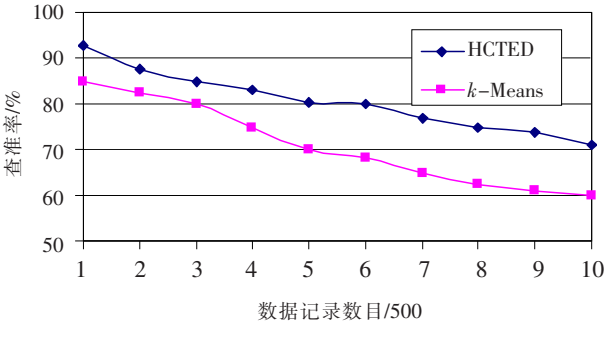

图 9 HCTED 与朴素算法运行查准率对比

Fig.9 Comparison of precision between HCTED and k-Means clustering algorithm

从图 9 和图 10 中可以发现,HCTED 算法的聚类准确性要明显高于 k-Means 算法,查准率一般在 70%以上,查全率一般在 60%以上。当数据记录较大

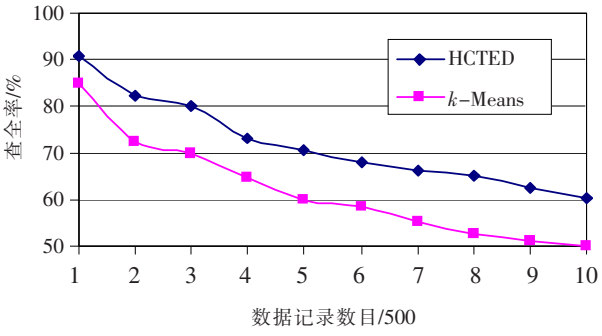

图 10 HCTED 与朴素算法运行查全率对比

Fig.10 Comparison of recall between HCTED and k-Means clustering algorithm

时,HCTED 准确性下降较慢，原因在于文中采用树编辑距离计算个体属性相似度准确率较高，此外与用于聚类的属性选取有关，如果用于聚类的属性选择不恰当将影响实验效果。

6.6 实验结论

通过上述实验,作者得出结论:利用基于树编辑距离的层次聚类划分算法进行身份识别是一种切实可行的有效方法。该算法的有效性在真实数据测试中得到验证,运行效率也得到充分保证。通过设置适当的参数,如阈值、队列窗口大小等可以在保证算法聚类准确率基础上对时间代价进一步优化，取得更好的实验效果。

7 结论及未来工作

文章借助树编辑距离这一概念来度量不同个体之间的相似度,并将其应用到聚类算法中。对属性采用层次编码方法，在给定阈值的前提下使用聚类算法对不同身份信息中具有相似特征的部分进行聚类操作,将其分类到不同集合中。基于该方法,可以有效辨识出身份信息中经过编造、篡改的虚假信息,进而辅助公安部门对犯罪分子进行准确的身份识别。通过实验对影响聚类过程的各种参数进行分析,并给出具体的优化方法。实验结果充分证实了基于树编辑距离的层次聚类算法 HCTED 在聚类准确性和效率上明显优于传统聚类算法，可有效地识别身份各分量信息，对提高身份鉴定工作效率有积极的意义。下一步工作包括:

(1)完善和优化 HCTED 算法,进一步提高算法的准确性和缩短运行时间;

(2)将层次编码方法应用到其他聚类算法中,如基于密度和网格的聚类算法;

(3)将 HCTED 算法应用到基于网络犯罪的识别和预警中，用于识别各种网络犯罪案件之间的相似性,辅助网络案件的侦破。

References:

[1] Liu Wei, Tang Changjie, Qiao Shaojie, et al. A new method for crime data mining based on conceptual e-mail system[J]. Computer Science, 2007,34(2):213-215.

[2] Qiao Shaojie, Tang Changjie, Peng Jing, et al. VCCM mining: mining virtual community core members based on gene expression programming[C]//LNCS 3917: WISI 2006, 2006:133-138.

[3] Wen Fenlian. The research of the key techniques in crime data mining[D]. Chengdu: Sichuan University, 2007.

[4] McAndrew D. The social psychology of crime: groups, teams, and networks [M]. Canter D, Alison L. UK: [s.n.], 2000:53-94.

[5] Xu J, Chen H. CrimeNet explorer: a framework for criminal network knowledge discovery[J]. ACM Transactions on Information Systems, 2005,23(2):202-226.

[6] Levenshtein V I. Binary codes capable of correcting deletions, insertions, and reversals[J]. Doklady Akademii Nauk SSSR, 1965,163(4):845-848.

[7] Tai K C. The tree-to-tree correction problem[J]. Journal of the Association for Computing Machinery, 1979,26(3): 422-433.

[8] Wagner R A, Fisher M J. The string-to-string correction problem[J]. Journal of the ACM, 1974,21(1):168-173.

[9] Ramsey J O. A PROC MATRIX program for preference-dissimilarity multidimensional scaling[J]. Psychometrika, 2006,51(1):163-170.

[10] Han Jiawei, Micheline K. Data ming: concepts and techniques[M]. New York: Morgan Kaufmann Publishers, 2000:238-243.

[11] Peng Jing, Tang Changjie, Cheng Wenquan, et al. A

hierarchy distance computing based clustering algorithm[J]. Chinese Journal of Computers, 2007,30(5):786-795.

[12] Braha D, Elovici Y, Last M. A theory of actionable data mining with application to semiconductor manufacturing control[J]. International Journal of Production Research, 2007,45(13):3059-3084.

附中文参考文献:

[1] 刘威,唐常杰,乔少杰,等.基于概念邮件系统的犯罪数据挖掘新方法[J].计算机应用,2007,34(2):213-215.

[3] 温粉莲.基于犯罪数据挖掘系统的关键技术研究[D].成都:四川大学,2007.

[11] 彭京,唐常杰,程温泉,等.一种基于层次距离计算的聚类算法[J].计算机学报,2007,30(5):786-795.

乔少杰(1981-),男,山东招远人,博士生,2004年于四川大学获计算机科学与技术专业学士学位,目前为四川大学计算机学院计算机应用专业博士生及新加坡国立大学计算机学院访问学生,主要研究领域为数据挖掘,数据库与知识系统。

QIAO Shaojie was born in 1981. He received his BS degree in Computer Science and Technology from Sichuan University in 2000. Now he is a PhD candidate at Sichuan University and also a visitor scholar in National University of Singapore, all in Computer Science. His research interests include data mining, database and knowledge system.

唐常杰(1946-),男,重庆人,教授,博士生导师,1982年于四川大学数学系获理学硕士学位,目前是四川大学计算机学院教授、博士生导师,主要研究领域为数据库,数据挖掘。

TANG Changjie was born in 1946. He received his MS degree in Mathematics from Sichuan University in 1982. He is currently a full professor and doctoral supervisor at Sichuan University. His research interests include database and data mining.

陈瑜(1974-),男,陕西汉中人,2005年于成都理工大学获计算机应用专业硕士学位,目前为四川大学计算机学院计算机应用专业博士生,主要研究领域为数据挖掘,数据库,基因表达式编程。

CHEN Yu was born in 1974. He received his MS degree in Computer Science from Chengdu University of Technology in 2005, and now is a PhD candidate in Computer Science at Sichuan University. His research interests include data mining, database and gene expression programming.

彭京(1973-),男,四川成都人,2005年于四川大学获计算机应用专业博士学位,目前为北京大学信息科学技术学院博士后,主要研究领域为数据挖掘,进化计算。

PENG Jing was born in 1973. He received his PhD degree in Computer Science from Sichuan University in 2005, and now is a Postdoctor in Computer Science at Peking University. His research interests include data mining and evolutionary computing.

温粉莲(1982-),女,云南曲靖人,2007年于四川大学获计算机科学与技术专业硕士学位,目前就职于西门子系统与软件工程(南京)有限公司,主要研究领域为数据挖掘,数据库。

WEN Fenlian was born in 1982. She received her MS degree in Computer Science from Sichuan University in 2007, and now is working at Siemens Program and System Engineering (Nanjing) Co., Ltd. Her research interests include data mining and database.