一种基于信任值的分类属性聚类算法*

李 梓,蒋庆丰,程晓旭,贾美娟
(大庆师范学院 计算机科学与技术学院,黑龙江 大庆 163712)

摘 要:针对 K-Modes 算法的不足,提出了一种基于信任值的分类属性聚类算法 TrustCCluster,该算法不需预先给定聚类个数,聚类结果稳定且不依赖于初始值的选取。在真实数据上验证了 TrustCCluster 聚类算法,并与 K-Modes 和 P-Modes 算法进行了对比,实验结果表明 TrustCCluster 算法是有效、可行的。

关键词:信任值;聚类;K-Modes 算法;P-Modes 算法
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2012)22-0057-03

A categorical attribute clustering algorithm based on trust value

Li Zi, Jiang Qingfeng, Cheng Xiaoxu, Jia Meijuan
(Department of Computer Science and Technology, Daqing Normal University, Daqing 163712, China)

Abstract: For the shortage of K-Modes algorithm, a categorical attribute clustering algorithm TrustCCluster based on trust value is proposed, the algorithm does not need to pre-specify the number of clusters, and clustering results do not depend on the selection of the initial values. TrustCCluster clustering algorithm is verified on the real data, and compared with the K-Mode and P-Modes algorithms, the result shows that TrustCCluster algorithm is feasible and effective.

Key words: trust value; cluster; K-Modes algorithm; P-Modes algorithm

聚类分析是一种非常重要的数据挖掘技术,已经被广泛应用于网络入侵检测、模式识别、图像处理、空间数据分析等领域。传统聚类算法可以分为基于划分的方法(如 K-Means)[1]、基于层次的方法(如 AGNES)[2]、基于密度的方法(如 DBSCAN)[3]、基于网格的方法(如 STING)[4]、基于模型的方法(如 SOM)[5]等。

K-Means 算法是一种基于划分的典型算法,具有描述容易、实现简单且快速等优点。但 K-Means 算法只能处理数值属性。为克服这一局限,Huang 等人提出了一种适合于处理分类属性数据的 K-Modes 算法[6-7]。该算法对于每个分类属性采用取值频度最大的属性值——模(mode)来表示类的对应“中心”,但这种表示难以准确反映类中对象的取值情况,导致距离计算不准确,并且当某个属性取值频度最大的属性值多于一个时,mode 值不唯一。P-Modes 算法[8]基于粗糙集理论,提出了一种新的距离度量方法,该距离度量在度量同一分类属性下两个属性值之间的差异时,克服了简单 0-1 匹配差异法的不足,既考虑了它们本身的异同,又考虑了其他相关分类属性对它们的区分性。

K-Modes、P-Modes 算法都是在 K-Means 算法基础上产生的一种针对分类属性的距离度量方法,和 K-Means 算法一样存在以下几点不足:(1) 需要预先给定聚类个数 K;(2)算法对初始值的选取依赖性极大;(3)算法容易陷入局部最优解。针对以上算法的不足,结合 P-Modes 的度量方法本文提出了一种基于信任值的分类属性聚类算法 TrustCCluster,该算法无需预先给定聚类个数,聚类结果稳定,不依赖于初始值的选取,具有更高的聚类精度。

1 基于信任值的聚类算法

1.1 基本概念

本文采用 P-Modes 中距离度量方法来表示具有分类属性的两个数据点之间的距离。设数据集 D 有 n 个元素,P 个分类属性。

定义 1 D 中任意两个数据 x_i、x_j 之间的距离为:

$$d(x_i,x_j)=1/P\sum_{k=1}^{k=P}d(x_{ik},x_{jk})$$

* 基金项目:黑龙江省自然科学基金项目(F200923);黑龙江省教育厅科学技术研究项目(11553001)

其中，$d(x_{ik},x_{jk})$为 x_i、x_j 的第 k 个属性值之间的距离。

定义 2 D 中任意两个数据 x_i、x_j 的第 k 个分类属性值之间的距离 $d(x_{ik},x_{jk})$为：

$$d(x_{ik},x_{jk})=1/P\sum_{l=1}^{l=P}d_l(x_{ik},x_{jk})$$

其中，$d_l(x_{ik},x_{jk})$为两个属性值 x_i、x_j 相对于属性 l 的距离。

定义 3 设数据集 D 中第 k 个属性的不同值的集合为 $T_k=\{T_{k1},T_{k2},\cdots,T_{kt}\}$，$1\leqslant t\leqslant|V_k|$，$|V_k|$为第 k 个属性不同值的个数。数据 x 的第 k 个属性值为 $x.V_k$。任意两个数据 x_i、x_j 的第 k 个分类属性值相对于属性 l 的距离 $d_l(x_{ik},x_{jk})$为：

$$d_l(x_{ik},x_{jk})=\begin{cases}1,x_{ik}\neq x_{jk}\\0,x_{ik}=x_{jk}\end{cases},k=l$$

$$d_l(x_{ik},x_{jk})=\begin{cases}0,x_{ik}=x_{jk}\\1/n\sum_{t=1}^{|V_l|}||C_t\cap X|-|C_t\cap Y||,x_{ik}\neq x_{jk}\end{cases},k\neq l$$

$X=\{x|x.V_k=x_{ik}\}$，$Y=\{x|x.V_k=x_{jk}\}$，即 X、Y 为其第 K 个属性值分别为 x_{ik}、x_{jk} 的数据集合。

$C_t=\{x|x.V_l=T_{lt}\}$，$1\leqslant t\leqslant|V_l|$。$T=\{T_{l1},T_{l2},\cdots,T_{lt}\cdots\}$。$C_t$ 是由第 l 个属性将数据集划分的等价类集合。

定义 4 信任值。一个数据点的信任值为以其为中心，半径为 R 的圆内包含的数据点的个数（不含自身）。

如图 1 所示，圆 1 的中心点所在圆内还有 4 个点，所以其信任值为 4，圆 2 的中心点信任值为 3，圆 3 的中心点信任值为 0，圆 1 和圆 2 相交点的信任值为 2，其他点信任值为 1。

图 1 信任值和共享信任值

定义 5 共享信任值。半径为 R，中心不同的两个相交圆的共享信任值为相交部分包含的数据点数。

如图 1 所示，圆 1 与圆 2 相交部分包含 1 个点，所以圆 1 和圆 2 的共享信任值为 1。

1.2 算法描述

TrustCCluster 算法具体步骤如下：

(1)计算信任值

所有数据点信任值初始值为 0，遍历数据集中所有数据点，若任意两个数据点 x_i、$x_j(1\leqslant i,j\leqslant n)$的距离 $d(x_i,x_j)<$半径阈值 R，则 x_i、x_j 的信任值都加 1，否则不变。

(2)按信任值大小对所有数据点进行降序排序

(3)聚类形成簇

①$i=1$(i 为数据点的序号)；

②若 x_i 未被访问过，则 x_i 作为新生成簇 Cluster 的中心，否则转步骤⑥；

③$j=i$；

④如果 $d(x_i,x_j)<R$ 且 x_j 未被访问过，则将 x_j 加到簇 Cluster 中并将 x_j 标记为已访问过；

⑤j 加 1，若 $j\leqslant n$，转步骤④；

⑥i 加 1，若 $i\leqslant n$，转步骤②；

⑦结束。

如图 1 所示，在半径阈值为 R 时，形成了 3 个簇，簇 1 的中心是信任值为 4 的点，一共包含 5 个点，簇 2 中心是信任值为 3 的点，包含 3 个点（不包括相交部分的点），簇 3 只包含 1 个点。

(4)合并簇

假设步骤(3)中，聚类形成 M 个簇。对于生成的 M 个簇，如果任意两个簇 $Cluster_i$、$Cluster_j$ 的共享信任值 $Share\geqslant(Cluster_i.sonNum+Cluster_j.sonNum)\times Q$，sonNum 为簇的成员数即簇包含的数据点数，$Q$ 为给定的比例阈值，$0<Q\leqslant 1$，则这两个簇合并形成新的簇。可知簇合并关系 R'是自反、对称、传递的，最终形成的簇是关系 R'对 M 个簇的一个等价划分。

如图 1 所示，如果 $Q\leqslant 1/(5+3)=0.125$，则簇 1 和簇 2 会合并成一个新的簇。

(5)输出聚类结果

算法代码如下：

```
//Step 1 计算信任值
for(i=1;i<n;i++)
    for(j=i+1;j≤n;j++)
    {
        if(d(xi,xj)<R)
        {
            xi.trust++;
            xj.trust++;
        }
    }
//Step 2 按信任值排序
    sort(dataSet);
//Step 3 聚类形成簇
for(i=1;i≤n;i++)
{
    if(visit[i] == false)
    {
        for(j=i;j≤n;j++)
    { //找出生成簇的成员
        //判断 xj 和簇 clusterNum 中心的距离是否<R
        if(! visit[j]&& d(clusterNum,xj)<R)
        {
            Cluster[clusterNum].sonNum++;
            visit[j]=true;
        }
    visit[i] = true;
}
```

```
    }
}
    // Step 4 合并簇
for(i=1; i≤clusterNum; i++)
    for(j=i+1; j≤clusterNum;j++)
{           //Matrix[M][M]为关联矩阵,所有元素初始值为 0
    if(Share>=(Cluster[i].sonNum+ Cluster[j].sonNum)*Q)
      {   //共享信任值 Share≥阀值,则关联矩阵的项为 1
    Matrix[i][j]=1;
    Matrix[j][i]=1;
  }
}
DFSMergeCluster();          //DFS 搜索求无向图连通分量,
                                即进行簇合并
```

1.3 算法分析

(1)TrustCCluster 算法需要设定两个参数:半径阈值 R 和比例阈值 Q。虽然 TrustCCluster 算法比 K-Modes 算法多了一个参数,却可以在预先不知道聚类个数 K 的情况下进行聚类。

(2)算法求出所有数据的信任值并排序,参数确定后,聚类结果也就确定,与初始数据点的选取顺序无关。

(3)信任值越大,表明数据点中心作用越大。算法求出所有数据的信任值并按信任值从大到小进行排序,聚类结果具有全局性。

(4)算法的步骤(3)中只能发现一些近似球形的小簇,但在步骤(4)中进行的簇合并操作,可形成非球形的大簇。

(5)时间复杂度分析。

算法首先根据定义 2、3 计算所有属性(假设第 k 个属性)下任意两个属性值之间的距离 $d(x_{ik},x_{jk})$,然后将其值存到数组 DisAttr[k][i][j]中,以便计算距离 $d(x_i,x_j)$ 时使用。计算距离 $d(x_i,x_j)$ 的时间复杂度为 $O(Psn+Ps^2n+P^2s^3n)=O((Ps+Ps^2+P^2s^3)n)$,$s=\max|V_l|$,$1\leqslant l\leqslant P$。由于对于大数据集 $n>>s$,$n>>P$,所以复杂度是线性 $O(n)$ 的。

步骤(1)中计算信任值的时间复杂度为 $O(n(n-1)/2)$,步骤(2)如果采用快速排序,则时间复杂度为 $O(n\lg n)$,步骤(3)聚类形成簇的时间复杂度为 $O(n(n+1)/2)$,步骤(4)合并簇的时间复杂度为 $O(m^2)$,m 为步骤(3)中生成的簇的个数($n>>m$)。

所以 TrustCCluster 算法的时间复杂度为 $O(n^2)$。

2 实验结果

为了验证本文算法的有效性,在 UCIMachine Learning Repository 提供的 Soybean 和 Congressional Vote 数据集上进行了测试,并与 K-Modes、P-Modes 算法进行了对比,结果表明本文算法聚精度更高。实验所用编程环境为 VC++6.0。

为比较聚类结果,定义聚类精度 r:

$$r=\sum_{i=1}^{k}a_i/n$$

其中,k 为聚类个数,a_i 为第 i 个聚类中占支配地位的类别的数据数,n 为总的数据个数。

Soybean 数据集表示大豆疾病的数据,共有 47 个数据,包括 35 个分类属性,并分为 4 类。

设 $E=\sum_{i=1}^{n}\sum_{j=i+1}^{j=n}d(x_i,x_j)/n\times(n-1)$ 为所有数据点的平均距离的一半,$T=\sum_{i=1}^{i=n}T_i/n^2$,$T_i$ 为数据 i 的信任值。TrustCCluster 在 $3/2E\leqslant R\leqslant 9/8E$,$31/16T\leqslant Q\leqslant 7/4T$ 时随机运行 100 次。K-Modes、P-Modes 算法的初始聚类个数为 $K=3$,初始点随机选取,运行 100 次。三种算法聚类精度比较如表 1 所示。可以看出 TrustCCluster 算法最大、最小、平均聚类精度都要高于其他两种算法,并且最大精度可达到 100%。

表 1 Soybean 数据集聚类结果

算法	最大聚类精度/%	最小聚类精度/%	平均聚类精度 r/%
K-Modes	85.11	57.45	75.30
P-Modes	94.77	73.54	89.45
TrustCCluster	100.00	78.72	93.94

Congressional Votes 数据集记录了美国国会 1984 年 435 个国会议员的投票情况。每条记录为一个议员在 16 个属性上的表决情况,并有一个分类标志 Republican 或 Democrat,用以表明对应的议员所属党派。每条记录由16 个属性描述,每个属性仅取 3 个值(y 表示同意,n 表示不同意,?表示不表态)。TrustCCluster 在 $4/5E\leqslant R\leqslant 5/4E$,$3/5T\leqslant Q\leqslant 3/2T$ 时随机运行 100 次。K-Modes、P-Modes 算法的初始聚类个数为 $K=3$,初始点随机选取,运行 100 次。三种算法聚类结果比较如表 2 所示。可以看出在适当参数下,TrustCCluster 算法最大聚类精度和平均聚类精度都要高于另外两种算法。

表 2 Congressional Votes 数据集聚类结果

算法	最大聚类精度/%	最小聚类精度/%	平均聚类精度 r/%
K-Modes	85.75	61.38	84.14
P-Modes	91.56	63.12	87.80
TrustCCluster	92.88	62.07	88.61

本文提出了一种基于信任值的分类属性聚类算法 TrustCCluster,相对于 K-Modes、P-Modes,TrustCCluster 算法具有如下优点:不需预先给定聚类个数;聚类结果不依赖于初始值的选取;可以发现非球形的簇;聚类精度更高。

TrustCCluster 算法的时间复杂度为 $O(n^2)$,不适合大规模数据的聚类,只支持分类属性数据。如何使 TrustCCluster 算法应用于大规模数据的聚类,并且能处理混合型数据,以及更合理地选择参数 R 和 Q 是需进一步研

究的问题。

参考文献

[1] MACQUEEN J.Some methods for classification and analysis of multivariate observations[C].Proc 5th Berkeley Symposium Mathematics Statist and Probaility, 1967:281-297.

[2] KAUFMAN J, ROUSSEEUW P J.Finding groups in data: an introduction to cluster analysis[M].New York: John Wiley&Sons, 1990.

[3] ESTER M, KRIEGEL H P, SANDER J, et al.A density-based algorithm for discovering clusters in large spatial databases[C].Proc.of 1996 Intl.Conf.on Knowledge Discovery and Data Mining, Portland, OR.1996:226-231.

[4] WANG W, YANG J, MUNTZ R.STING: A statistical information grid approach to spatial data mining[C].Proc of 1997 Intl.Conf.on Very Large Databases, Athens, Greece. 1997:186-195.

[5] KOHONEN T.Self-organized formation of topologically correct feature maps[J].Biological Cybernetics, 1982, 43(1): 59-69.

[6] Huang Zhexue.Clustering large data sets with mixed numeric and categorical values[C].Proc of PAKDD 97.Singapore: World Scientific, 1997:21-35.

[7] Huang Zhexue.Extensions to the K-means algorithm for clustering large data sets with categorical values[J].Data Mining and Knowledge Discovery, 1998, 2(3):283-304.

[8] 梁吉业,白亮,曹付元.基于新的距离度量的K-Modes聚类算法[J].计算机研究与发展, 2010, 47(10):1749-1755.

(收稿日期:2012-07-24)

作者简介:

李梓,女,1970年生,硕士研究生,副教授,主要研究方向:数据挖掘,网络安全。

蒋庆丰,男,1983年生,硕士研究生,讲师,主要研究方向:移动计算,可信计算。

程晓旭,女,1965年生,硕士研究生,副教授,主要研究方向:网络安全。