

一种有效的高维分类数据聚类方法研究

贾俊芳^{1,2,3}, 李德玉^{1,2}

(1 山西大学 计算机与信息技术学院, 山西 太原 030006;

2 山西大学 计算智能与中文信息处理教育部重点实验室, 山西 太原 030006;

3 山西大同大学 数学与计算机科学学院, 山西 大同 037006)

摘 要: 随着数据规模的不断增大, 提高 K -modes 聚类算法或模糊 K -modes 聚类算法的运行效率成为了一个重要问题. 为了提高其算法执行效率, 提出了一种基于分治法的高维分类数据聚类方法. 该方法并不是一次性对所有的数据进行聚类, 而是将分类数据集分成若干个子集, 对每个子集同时进行聚类, 最后对聚类结果进行融合以形成最终的聚类结果. 实验结果表明大多数情况下较传统的方法在聚类的速度上有显著的提高.

关键词: 聚类分析; 模糊聚类; 分治法; 分类数据; 评价指标

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2011)06-0088-04

An Effective High Dimensional Categorical Data Clustering Method Research

JIA Jun-fang^{1,2,3}, LI De-yu^{1,2}

(1 School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China;

2 Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China; 3 College of Methmatic and Computer Technology, Datong 037009, China)

Abstract: With the increasing size of data set, improving the efficiency of K -modes clustering algorithm or fuzzy K -modes clustering algorithm is becoming a critical issue. In order to improve the efficiency of the algorithm, a clustering method based on divided and conquered method was proposed. This method, not a one-time clustering of all data, divided the data set into several subsets, and each subset was clustered at the same time; the fusion results of each subset cluster form the final clustering results. The results show that the efficiency of clustering has been increased greatly compared with traditional clustering method in most cases.

Key words: clustering analysis; fuzzy clustering; divided and conquered method; large categorical data sets; evaluation index

1 引言

聚类在模式识别、数据挖掘、机器学习、图像处理等领域发挥着重要作用. 所谓的聚类就是根据若干个属性, 按照一定的相似性或不相似性度量, 把数据集分成若干类, 其目的是要让同类的数据间的相似度尽量大, 不同类的数据间相似度尽量小. 文献

[1-2]中对聚类算法的各个方面进行了总结. 聚类算法基本上可分为层次聚类、划分聚类、基于密度的聚类、基于模型的聚类和基于网格的聚类. 其中, 划分聚类中的 K -均值聚类算法应用广泛. 但是传统聚类大都针对连续的数值型数据. 如果将它们直接应用到分类数据中, 效果往往不尽人意. 而现实中, 分类数据常常出现.

收稿日期: 2010-11-15; **修回日期:** 2011-01-30

基金项目: 国家自然科学基金资助项目(60970014); 教育部高等学校博士点基金(200801080006); 教育部科学技术研究重点项目(207018); 山西省重点实验室开放基金项目(2007031017); 太原市科技明星专项基金项目(09121001)

已有的分类数据聚类算法有 K -modes 聚类算法和模糊 K -modes 聚类算法^[4]. 然而,这些聚类算法的时间消耗将非常大,面对海量的高维符号数据时这些时间消耗将不可容忍. 因此,提高效率显得尤为重要. 鉴于此,本文提出基于分治法^[7]的分类数据聚类分析的有效方法对原有算法在聚类速度上进行一定的改进.

2 模糊 K -modes(FKMD)聚类算法

模糊 K -modes 聚类算法是在 K -modes 的算法^[3]的基础上,对模糊 K -modes 算法的扩展. 它用 modes 代替了 means,并且对每个聚类中的对象分配隶属度,使得对于具有分类属性值的数据能够得到较好的聚类结果. 算法类似于模糊 K -modes 算法,交替更新聚类中心和划分矩阵,直到代价函数值不再变化. 其简单描述如下:

假定 $X = \{X_1, X_2, \dots, X_n\}$ 是一组数据元组,其中 $X_i = [x_{i1}, x_{i2}, \dots, x_{im}]$ 表示具有 m 个属性的数据对象. 设 k 是一个正整数,将 X 中的 n 个对象划分到 k 个的类别中. 用使以下代价函数最小作为聚类的准则:

$$E_c(W, Q) = \sum_{l=1}^k \sum_{i=1}^n \omega_{li}^a d_c(Q_l, X_i) \quad (1)$$

这里, $Q_l = [q_{l1}, q_{l2}, \dots, q_{lm}]$ 是能够代表聚类 l 的向量,即 modes, $\omega_{li} \in [0, 1]$ 是划分矩阵 $W_{l \times n}$ 的一个元素,它表示对象 i 划分到聚类 l 中的隶属度, $\sum_{i=1}^n \omega_{li} = 1$, d_c 是差异测度, $a > 1$ 是加权指数. 对于分类数据,定义差异测度:

$$d_c(Q_l, X_i) = \sum_{j=1}^m \delta(q_{lj}^c, x_{ij}^c) \quad (2)$$

式中,当 $p = q$ 时, $\delta(p, q) = 0$, 否则, $\delta(p, q) = 1$. x_{ij}^c 和 q_{lj}^c 是在第 j 个分类属性上的取值, m 是属性的个数.

划分矩阵的更新方法如下:

$$\omega_{li} = 1 \quad \text{if} \quad X_i = Q_l$$

$$\omega_{li} = 0 \quad \text{if} \quad X_i = Q_h, h \neq l$$

$$\omega_{li} = \frac{1}{\sum_{h=1}^k \left[\frac{d_c(Q_l, X_i)}{d_c(Q_h, X_i)} \right]^{1/(a-1)}} \quad \text{if} \quad X_i \neq Q_l, X_i \neq Q_h, h = 1, 2, 3, \dots, k.$$

聚类中心的更新方法如下: 设 $A_j = \{a_j^{(1)}, a_j^{(2)}, \dots, a_j^{(n_j)}\}$, 其中 n_j 是第 j 个分类属性的可取值的个数, 聚类中心 $Q_l = [q_{l1}, q_{l2}, \dots, q_{lm}]$, $l = 1, 2, \dots, k$.

有如下定理: 式(1)取的最小, 当且仅当 $q_{lj} =$

$a_j^{(r)} \in A_j, j = 1, 2, \dots, m$, 其中 $a_j^{(r)}$ 满足下式:

$$\sum_{i, x_{ij}=a_j^{(r)}} \omega_{li}^a \geq \sum_{i, x_{ij}=a_j^{(t)}} \omega_{li}^a, 1 \leq t \leq n_j \quad (3)$$

即, $a_j^{(r)}$ 是在属性的不同数值的数据集上, 分别求其属于聚类 l 的隶属度的幂和和其中最大幂所对应的属性值.

3 分治思想

将一个规模为 n 的问题分解为 k 个规模较小的子问题, 这些子问题相互独立且与原问题相同. 递归地解这些子问题, 然后将各子问题的解合并得到原问题的解. 它的一般算法设计模式如下:

```

Divide-and-Conquer( $n$ )
{
  if ( $n \leq n_0$ )
    {解子问题;
     return(子问题的解);}
  for ( $i = 1; i \leq k; i = i + 1$ )
    {分解原问题为更小的子问题  $P_i$ ;
      $y_i = \text{Divide-and-conquer}(|P_i|)$ ;
      $T = \text{MERGE}(y_1, y_2, \dots, y_k)$ ;
     return( $T$ );}
}

```

其中 $|P|$ 表示问题 P 的规模; n_0 为一阈值, 表示当问题 P 的规模不超过 n_0 时, 问题已容易直接解出, 不必再继续分解. 算法 $\text{merge}(y_1, y_2, \dots, y_k)$ 是该分治法中的合并子算法, 用于将 P 的子问题 $P_1, P_2, \dots, P_k$ 的相应的解 $y_1, y_2, \dots, y_k$ 合并为 P 的解.

4 基于分治法的分类数据聚类算法

4.1 算法描述及流程图

算法过程描述如下:

- (1) 给定数据集 $A = \{a_1, a_2, \dots, a_n\}$;
- (2) 将数据集 A 随机的分成若干个子集 $A_1, A_2, \dots, A_p$;
- (3) 对子集 $A_1, A_2, \dots, A_p$ 同时进行聚类, 分别得到的聚类中心数为 $m_1, m_2, \dots, m_p$;
- (4) if $|m_1 + m_2 + \dots + m_p| \geq n_0$ (n_0 为问题规模的阈值), then 把 $m_1, m_2, \dots, m_p$ 个聚类中心看成集合 A , 转到(2); else 转到(5);
- (5) 把 $m_1 + m_2 + \dots + m_p$ 个数据进行一次性聚类;

(6) 类的合并: 第(5)步结束后, if 聚类中心 x_1 和 x_2 聚为一类, 而在第(3)步结束后 c_1 和 c_2 分别是以 x_1 和 x_2 为聚类中心的类, then 就把类 c_1 和 c_2 合并为一类.

4.2 性能分析

本文提出的算法能有效地提高大数据集的聚类速度. 首先与传统的聚类算法 K -modes 作比较, 在 K -modes 聚类算法中每一次迭代中需要执行的步骤如图 1 所示.

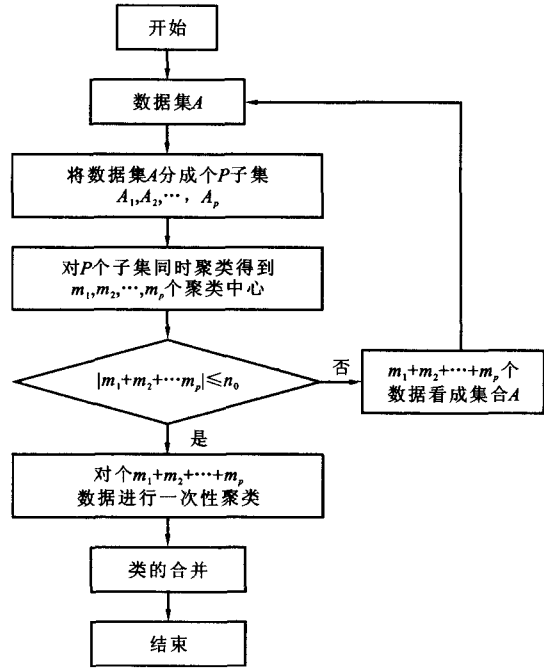


图 1 算法实现流程图

(1) 对于每一个 $h \neq l$, 如果 $d(z_l, x_i) \leq d(z_h, x_i)$ 则 $w_{li} = 1$, 否则 $w_{li} = 0, i = 1, 2, \dots, n, 1 \leq h \leq k$.

(2) $f_r(A_j = q_j | X) \geq f_r(A_j = c_{k,j} | X)$ for $q_j \neq c_{k,j}$ for all $j = 1, 2, \dots, m$.

如果一共有 n 个数据, 且分为 k 类, 那么每一次迭代在第一步中需要比较 $n \times k$ 次, 若需要迭代 L 次, 那么一共需要比较 $L \times n \times k$ 次, 在第二步中需要计算 $k \times L$ 次, 所以对 n 个数据聚类一共需要的时间 $T(n) = o(L \times k \times n + k \times L)$. 由于 n 很大, 所以 $L \times k \times n$ 会很大, 这样执行的效率会非常低. 而本文提出的方法采用分治法, 如果一个分治法将规模为 n 的问题分成 s 个规模为 n/m 的子问题来解决, 那么可得到执行时间 $t(n) = o(n^{\log_m^s})$. 如果将 n 个数据平均分为 s 个子集, 那么每个子集就有 n/k 个数据, 此时 $s = m$, 那么 $t(n) = o(n)$, 由于 $o(n) < o(L \times k \times n + k \times L)$, 可见此时的时间复杂度降低了很多, 极大地提高了聚类的速度.

然后再与 FKMD 算法作比较, FKMD 算法聚类这 n 个数据分为 c 类所要的时间为 $T(n) = o(L \times c \times n + L \times c)$, 而由于 n 很大, 所以执行速度很慢. 而

采用本文提出的方法, 如果我们把数据 n 平均分成 k 个子集, 那么每个子集的数据量为 n/k 个, 同样可得到此时的执行时间为 $t(n) = o(n)$, 而 $o(n) < o(L \times c \times n + L \times c)$, 所以降低了时间复杂度, 极大地提高了效率.

5 实验结果

5.1 实验数据

为了证明该方法的有效性, 我们选取了 UCI 数据集上的 soybean、zoo、breast-cancer、mushroom 四个数据集. soybean 数据集包含 47 个数据对象, 每个数据对象有 35 个属性. zoo 数据集数据集包含 101 个数据对象, 每个数据对象有 16 个属性. breast-cancer 数据集包含 699 个数据对象, 每个数据对象有 9 个属性. mushroom 数据集包含 8 123 个数据对象, 每个数据对象有 22 个属性.

5.2 评价指标

用 AC(accuracy)、PE(precision)、RE(recall)^[4]

对实验结果进行评价. 在此 $AC = \frac{\sum_{l=1}^k a_l}{n}, PR = \frac{\sum_{l=1}^k (\frac{a_l}{a_l + b_l})}{n}$ and $RE = \frac{\sum_{l=1}^k (\frac{a_l}{a_l + c_l})}{n}$

K 为类的中心数, a_l 分类正确的对象的个数, b_l 分类不正确的对象的个数, c_l 本来属于某类但不在该类的对象的个数.

5.3 soybean 数据实验过程

(1) 对 soybean 数据数据用传统的 K -modes 算法一次性聚类, 程序执行 50 次求其 AC、PE、RE 的平均值及运行时间.

(2) 用分治法对 soybean 数据聚类, 通过迭代划分为两子集, 前 20 个数据划为一个子集 A , 后 21~47 个数据划分为一个子集 B . 分别用算法对两子集同时聚类, 设置每个子集的聚类中心数为 4, 程序各执行 50 次. 子集 A 聚类后的结果如表 1 所示, 子集 B 聚类后的结果见表 2.

表 1 对集合 A 聚类后的结果

聚类中心	属于该类的数据
500201210210030002100	$X_{12}, X_{13}, X_{17}, X_{18}, X_{19}, X_{20}$
302102021210301100000	X_2, X_3, X_5, X_7, X_8
602101010010311100000	$X_1, X_4, X_6, X_9, X_{10}$
600110210010030002100	$X_{11}, X_{14}, X_{15}, X_{16}$

表 2 对集合 B 聚类后的结果

聚类中心	属于该类的数据
012000110100110100030	$X_{21}, X_{22}, X_{23}, X_{24}, X_{25}, X_{26}, X_{41},$ X_{46}, X_{29}, X_{30}
112002121010220000031	$X_{31}, X_{33}, X_{35}, X_{36}, X_{38}, X_{39}, X_{40},$ X_{43}, X_{44}, X_{27}
011101110010120000031	X_{32}, X_{34}, X_{45}
012103110010120000031	$X_{37}, X_{42}, X_{47}, X_{28}$

(3) 将两个子集聚类后的聚类中心放在一起形成一个集合 C. 由于集合 C 已包含较少的数据,因此没有必要去递归,所以对集合 C 进行聚类,设置聚类中心数为 4. 最后进行类的合并. 结果见表 3.

从表 3 我们可以得到结果与一次性聚类结果基本一致. 我们利用该方法连续做实验 50 次,最后求出每次聚类结果的 AC、PR、RE 的平均值及运行时

间. 用同样的方法对 zoo、breast-cancer、mushroom 数据集进行聚类,求出其 AC、PR、RE 的平均值及运行时间. 得到表 4. 从表 4 可以看到随着数据量的增大,分治法的聚类效率显著的提高,聚类效果和一次性聚类的结果基本一致.

表 3 对集合 C 聚类后的结果并类进行合并

聚类中心	属于该类的数据
500201210210030002100	$X_{12}, X_{13}, X_{17}, X_{18}, X_{19}, X_{20}, X_{11}, X_{14},$ X_{15}, X_{16}
302102021210301100000	$X_2, X_3, X_5, X_7, X_8, X_1, X_4, X_6,$ X_9, X_{10}
012000110100110100030	$X_{21}, X_{22}, X_{23}, X_{24}, X_{25}, X_{26}, X_{41}, X_{46},$ X_{29}, X_{30}
	$X_{31}, X_{33}, X_{35}, X_{36}, X_{38}, X_{39}, X_{40}, X_{43},$
011101110010120000031	$X_{44}, X_{27}, X_{32}, X_{34}, X_{41}, X_{45}, X_{37}, X_{42},$ X_{47}, X_{28}

表 4 两种聚类方法实验得出的结果

	K-mode 算法				分治法			
	soybean	zoo	breast-cancer	mushroom	soybean	zoo	breast-cancer	mushroom
AC	0. 856 8	0. 835 7	0. 846 1	0. 731 2	0. 835 5	0. 829 1	0. 821 1	0. 711 1
PR	0. 900 8	0. 818 4	0. 870 2	0. 752 1	0. 890 0	0. 803 3	0. 856 3	0. 736 6
RE	0. 844 1	0. 617 0	0. 783 2	0. 726 9	0. 834 2	0. 700 0	0. 751 1	0. 701 2
Time/s	0. 063 2	0. 090 5	0. 180 3	2. 154 7	0. 042 6	0. 058 8	0. 112 1	1. 209 1

6 结束语

本文提出了一种基于分治法的分类数据集聚类方法,首先通过理论分析可知利用该方法能提高对分类数据集聚类的速度. 因为把分类数据集分成几个子集后,每个子集包括了较少的数据,这些子集可以并行地进行聚类,通过对各个聚类结果进行融合可以得到最后的聚类结果. 实验分析可得:利用该方法在大多数情况下与一次性聚类的结果基本一致的前提下可以大大节约算法运行时间.

参考文献:

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008(1): 48—61.

[2] Jain A K, Murty M N, Flynn P J. Data clustering: a review [J]. ACM Computing Surveys, 1999, 31(3):274—289.

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

[4] Michael K Ng, Mark Junjie Li, Joshua Zhexue Huang,

et al. On the impact of dissimilarity measure in k—modes clustering algorithm[J]. IEEE Transactions On Pattern Analysis and Machine Intelligence, 2007,29(3): 503—507.

[5] 吕国英,任瑞征,钱宇华. 算法设计与分析[M]. 2 版. 北京:清华大学出版社,2009:139—150.

[6] 蔡自兴,徐光祐. 人工智能及其应用[M]. 3 版. 北京:清华大学出版社,2006.

[7] Fuyuan Cao, Jiye Liang, Liang Bai. A new initialization method for categorical data clustering [J]. Expert Systems with Applications. 2009(36):10223—10228.

[8] 白亮,梁吉业,曹付元. 基于粗糙集的改进 K—mode 聚类算法[J]. 计算机科学, 2009, 36 (1) 162—164.

[9] 白亮,曹付元,梁吉业. 基于新的相异度量的模糊 K—modes 聚类算法[J]. 计算机工程 2009, 35(16):192—194.

作者简介:

贾俊芳 女,(1976—),硕士研究生. 研究方向为数据挖掘和人工智能.

李德玉 男,(1965—),教授,博导. 研究方向为计算智能与数据挖掘.