模糊聚类计算的最佳算法*

马 军[1], 邵 陆[2]

[1](山东大学 计算机科学系,山东 济南 250100);

[2](山东省医药工业研究所,山东 济南 250100)

E-mail: majun@cs.sdu.edu.cn

http://www.cs.sdu.edu.cn

摘要: 给出模糊关系传递闭包在对应模糊图上的几何意义,并提出一个基于图连通分支计算的模糊聚类最佳算法.对任给的 n 个样本,新算法最坏情况下的时间复杂性函数 $T(n)$ 满足 $O(n)\leqslant T(n)\leqslant O(n^2)$.与经典的基于模糊传递闭包计算的模糊聚类算法的 $O(n^3\log n)$ 计算时间相比,新算法至少降低了 $O(n\log n)$ 时间因子.理论分析与计算机实验表明,新算法对大规模数据进行模糊聚类计算的实际计算时间,在实际应用中是可以被接受的.

关键词: 模糊理论;模糊关系;模糊聚类;模糊应用

中图法分类号: TP181 **文献标识码**: A

模糊聚类(分类)分析技术是采用模糊数学方法,根据对象的各种属性(或因素),按某些预定指标进行分类的一门多元技术.在模糊聚类分析中,要进行分类的对象称为样本.设有 n 个样本 $U=(x_1,x_2,\ldots,x_n)$,其中每一个样本 x_i 具有 m 个特性指标,即 x_i 由向量 $x_i=(x_{i1},x_{i2},\ldots,x_{im})$ 表示.我们可用多元分析方法建立起样本之间的模糊关系 $R:U\times U\to[0,1]$,$(x_i,x_j)\in R$ 的程度通常用隶属函数 $\mu_R(x_i,x_j)$ 表示.$\mu_R(x_i,x_j)$ 被称为 x_i 与 x_j 的相似值,满足 $0\leqslant\mu_R(x_i,x_j)\leqslant 1$,$\mu_R(x_i,x_j)=\mu_R(x_j,x_i)$.一般来说,$\mu_R(x_i,x_j)$ 的值越大,两者的相似程度越高.定义 $\mu_R(x_i,x_j)=1$,$i=1,2,\ldots,n$.模糊关系的具体确定方法有距离法,如切比雪夫距离:$\mu_R(x_i,x_j)=1-c\max\limits_{1\leqslant i,j\leqslant m}|x_{ik}-x_{jk}|$,这里,$c$ 表示一个常数.另外,还有海明距离、欧氏距离以及相似系数法、贴近度法等,更详细的内容请参见文献[1~4].由 n 个向量之间的模糊关系形成了一个 $n\times n$ 关系矩阵 $R=(\mu_R(x_i,x_j))_{n\times n}$,称 R 为模糊相似关系矩阵[3].

设 U,V 和 W 分别表示 3 个论域,若 R,S 分别为 $U\times V$ 和 $V\times W$ 上的模糊关系,定义关系的合成 $R\circ S$ 为 $U\times W$ 上的模糊关系,其隶属函数定义为

$$\forall\, x\in U,\forall\, z\in W,\quad \mu_{R\circ S}(x,z)=\{\max_{y\in V}[\min(\mu_R(x,y),\mu_S(y,z)]\}. \tag{1}$$

若 R 是 U 上的模糊关系,并满足 $R\circ R\subseteq R$,则称 R 为模糊传递矩阵.

若我们所建立的模糊相似矩阵 R 满足传递性,则称 R 为等价关系.对任意 $\alpha\in[0,1]$,集合 $R\alpha=\{(x,y)\mid\mu_R(x,y)\geqslant\alpha\}$ 被称为 R 的 α-截集,α 称为截集阈值.若 R 为模糊等价关系,则对任意 $\alpha\in[0,1]$,$R\alpha$ 也为模糊等价关系[3].有时称 $R\alpha$ 为 α-截矩阵.若 $R\alpha$ 为一等价矩阵,则 $\forall\, x\in U$,$[x]=\{y\mid\mu_R(x,y)\geqslant\alpha\}$ 构成了 U 相对于阈值 α 的模糊聚类.目前,进行模糊聚类计算的最基本算法为下面的基于模糊矩阵传递闭包的方法[1~4].该算法可表述为

---

* 收稿日期: 1999-09-14; 修改日期: 1999-12-08

基金项目: 国家 863 高科技发展计划资助项目(863-306-ZT06-01-4);山东省自然科学基金资助项目(Z99G01)

作者简介: 马军(1956-),男,山东汶上人,博士,教授,主要研究领域为人工智能,算法分析与设计;邵陆(1959-),女,山东禹城人,高级工程师,主要研究领域为天然药物提取、分析与新药研制.

算法 1. 基于模糊关系传递闭包的基本模糊聚类算法

1. 确定 n 个样本 $X=(x_1,x_2,\ldots,x_n)$ 上的模糊相似关系 R 和一个截集阈值 α;
2. 将 R 按下面计算改造为一个等价矩阵

   $R\circ R=R^2$ //$R\circ R$ 为模糊关系的合成运算//

   $R^2\circ R^2=R^4$

   ...

   $R^{2^k}\circ R^{2^k}=R^{2^{k+1}}$

   直到存在一个 k,满足 $R^{2^k}=R^{2^{k+1}}$. //在 R 是相似矩阵的假设下,已证明必有这样的 k 存在,满足 $k\leqslant\log n$[3]. 以下用 R^* 表示 R 的传递闭包//
3. 计算集合 $[x]=\{y|R^*(x,y)\geqslant\alpha\}$. $[x]$即为模糊聚类.
4. 停止.

因为一次矩阵乘法所需时间为 $O(n^3)$,故上述算法的时间复杂性为 $O(n^3\log n)$. 这对早期的应用影响似乎不大,例如,按某些化学元素的含量对煤层、油层、钻石等进行等级分类,以及在日常生活、林业、矿藏开采等领域的应用等[1~6]. 因为上述问题一般可假设样本个数 $n<100$,所以此时微机的计算时间只需几秒. 而近年来,人们开始研究对数据库内的数据再利用,模糊聚类方法已被应用到数据挖掘、模式识别、机器概念学习以及决策支持等领域[3,5,6]. 例如,根据人口数据库,依据指标"体重/身高"及阈值 α,可把人分成"胖人集"、"不胖不瘦集"、"瘦人集"等概念聚类;对超市可根据"月底销售数量/月初库存"及阈值 α,确定出"畅销品集"与"滞销品集",并进一步分析在同一聚类集合内的商品之间的销售相关性、支持营销决策等. 模糊聚类技术还可应用到对多维图像的模式识别、天然药草按药性分类等[6]. 对于上述应用,一般样本数 n 变得相当大. 而当 $n>10^4$ 时,上述算法在微机上的计算时间则需几十天以上. 这使得聚类分析技术难以实际应用到大规模的数据分析中.

本文提出了一个全新的计算模糊聚类的最佳算法,其时间复杂性 $T(n)$ 满足 $O(n)\leqslant T(n)\leqslant O(n^2)$. 若把计算 $\min\{a,b\}$,$\max\{a,b\}$作为一个标准操作,以一台每秒可执行一百万次这样的标准操作的计算机为例,当 $n=10^4$ 时,原算法的计算时间需 80 多天,而新算法只需 6.7 秒左右. 计算机对应用实例的试验结果也与上述理论推算基本一致(依据计算机的配置,仅相差常数因子). 新算法的时间复杂性确保了对实际较大规模的数据进行模糊聚类计算的时间有效性.

1 新算法描述及理论分析

因为无向图可以表示关系,我们用无向赋权图 $G(V,E)$表示上述的相似矩阵 R,满足:若边 $e=(u,v)\in E$,当且仅当 $R(u,v)>0$. 通常 G 被称为模糊图[3]. 边 e 的权用 $C(e)$表示,并定义 $C(e)=R(u,v)$. 设 $P=v_{i_1},v_{i_2},\ldots,v_{i_j}$ 为 G 中的一条连接顶点 v_{i_1},v_{i_j} 的路,定义 P 的路宽 $\mathrm{wide}(P)=\min_{1\leqslant k\leqslant j}\{C(v_{i_k},v_{i_{k+1}})\}$. 设 $W(i,j)$为顶点 i,j 之间最宽路的路宽,基于 Floyd 全源最短路算法原理[7],可类推出 $W(=W_{ij}^{(n)})$,并可由以下递推公式计算:

$$W_{ij}^{(0)}=R(i,j);\quad 0\leqslant i,\ j\leqslant n; \tag{2}$$

$$W_{ij}^{(k)}=\max\{W_{ij}^{(k-1)},\min[W_{ik}^{(k-1)},W_{kj}^{(k-1)}]\};\quad 0\leqslant i,j\leqslant n;0\leqslant k\leqslant n. \tag{3}$$

若将式(1)和算法 1 中对 R^* 的计算过程与上述对 W 的计算过程相比较,不难发现二者的计算结果实质上是相同的,即模糊传递闭包 $R^*(u,v)$的几何意义为对应模糊图 G 中连结顶点 u 与 v 最宽路径的宽度值. 以此为桥梁,下述引理又进一步说明了模糊聚类与模糊子图连通分支之间的关系.

引理 1. 设 $G_\alpha=(V,E')$ 为模糊图 $G(V,E)$ 的一个子图,满足 $E'=\{(u,v)\mid u,v\in V$,并且 $R(u,v)\geqslant\alpha\}$,则 $R^*(u,v)\geqslant\alpha$ 的充要条件为 u,v 在 G_α 中,并在同一连通分支中.

证明:若 $R^*(u,v)\geqslant\alpha$,则有 $W(u,v)=R^*(u,v)\geqslant\alpha$. 即在模糊图中至少存在一条路径 P,满足 $\mathrm{Wide}(P)\geqslant\alpha$. 根据路宽的定义,可推出 P 中的边权均大于 α. 所以又推出 P 也应是 G_α 中的一条路径,故 u,v 在 G_α 中必在同一连通分支中.

假设 u,v 在 G_α 的同一连通分支中,则 G_α 中必存在连接两顶点的路径 P,满足 $\mathrm{wide}(P)\geqslant\alpha$. 因为 G_α 为 G 的一个子图,所以 P 也为模糊图 G 中的路径,故推出 $W(u,v)\geqslant\mathrm{wide}(P)\geqslant\alpha$. 这意味着 $R^*(u,v)=W(u,v)\geqslant\alpha$. □

引理 1 说明模糊传递闭包计算可以用对应于模糊子图 G_α 的连通分支计算代替. 考虑到在实际应用中,截集 α 的取值一般为几个固定值,故可认为 α 的取值个数为常数. 在此假设下,我们可以得到下述计算模糊聚类的新算法.

算法 2. 基于图的连通分支计算的模糊聚类算法

1. 确定 n 个样本 $X=(x_1,x_2,\ldots,x_n)$ 上的模糊相似关系 R 和一个截集阈值集合 S;
2. 对每个 $\alpha\in S$ 做
   2.1. 建立无向图 $G_\alpha(V,E)$;其中 V 对应 n 个样本;边 $(u,v)\in E$,当且仅当 $R(u,v)\geqslant\alpha$;
   2.2. 调用计算无向图的连通分支算法[8]计算 $G_\alpha(V,E)$ 的连通分支. $G_\alpha(V,E)$ 中惟一的连通分支内的顶点集合构成了所求模糊聚类集合.

定理 1. 对任意模糊关系与阈值 α,算法 2 正确地计算了关于 α 的模糊聚类,其时间复杂性为 $T(n)$,满足 $n\leqslant T(n)\leqslant n^2$. 该算法为计算模糊聚类的最佳算法.

证明:算法 2 的正确性已由引理 1 给出. 下面我们分析算法的时间复杂性. 因为在步骤 1 中,n 个样本 $X=(x_1,x_2,\ldots,x_n)$ 上的模糊相似关系 R 可由 m 个二元组表示,故步骤 1 的时间复杂性为 $O(m)=O(|E|)$.

在步骤 2 的每次循环中,无向图 G_α 可由图的边邻接表中 $2m$ 条边表示[8],所以步骤 2.1 的时间复杂性为 $O(|E|)$. 又因为基于边邻接表来计算图连通分支算法的时间复杂性为 $O(|E|)$[8],故算法中步骤 2.2 的时间复杂性也为 $O(|E|)$. 综上所述,步骤 2 的一次循环的执行时间为 $O(|E|)$. 因为假设阈值集合 S 内的阈值个数有限,故循环次数为常数,所以推出步骤 2 的时间复杂性为 $O(|E|)$.

基于对步骤 1 和步骤 2 的分析,推知新算法的时间复杂性 $T(n)=O(|E|)$,满足 $n\leqslant|E|\leqslant n^2$.

显然,在计算模糊聚类时,必须考虑所有 $R(u,v)\geqslant\alpha$ 的关系. 这等价于计算模糊聚类算法的时间下界函数 $\Omega(n)$ 应满足 $\Omega(n)\geqslant O(|E|)=T(n)$. 又因 $T(n)$ 为一个具体计算模糊聚类算法的执行时间,即 $T(n)$ 为计算模糊聚类的时间上界,故有 $T(n)=O(|E|)\geqslant\Omega(n)$. 综上两点,有 $T(n)=\Omega(n)$,即算法为计算模糊聚类的最佳算法. □

2 结束语

把模糊聚类分析技术有效地应用在大规模数据分析中是一个在数据挖掘、模式识别、决策支持应用领域中所遇到的实际问题,具有重要的理论与实际应用价值. 本文通过提出模糊图中路宽的概念,首次指出模糊传递闭包在相应模糊图上的几何意义,并以此为桥梁,提出利用图的连通分支算法来计算模糊聚类的新思路,这使得算法在计算时间复杂性上有了本质上的改进. 新算法不仅可以保证对大规模数据的模糊聚类计算的实时性,而且从计算复杂性的角度还说明了模糊聚类计算应属于实际容易计算的问题类.

致谢 本文作者衷心感谢审稿老师细心地指出原稿中的不足之处,并提出许多建设性的修改意见,使得本稿在组织及讨论的深度上有了很大的改善.

References:

[1] Wang, Pei-zhuang, Li, Hong-xing. Fuzzy System Theory and Fuzzy Computers. Beijing: Science Press, 1996. 166~191 (in Chinese).

[2] He, Zhong-xiong. Fuzzy Mathematics and Applications. Tianjin: Tianjin Science and Technique Press, 1984. 76~185 (in Chinese).

[3] He, Xin-gui. Fuzzy Theories and Fuzzy Techniques in Knowledge Processing. 2nd ed. Beijing: National Defence Industry Press, 1998. 414~421 (in Chinese).

[4] Zhang, Yue. Fuzzy Math Methods and Applications. Beijing: Coal Industry Press, 1992. 273~333 (in Chinese).

[5] Chen, Wen-wei. Intelligent Decision Techniques. Beijing: Publishing House of Electronics Industry, 1998. 3~21 (in Chinese).

[6] Shen, Qing, Tang, Lin. An Introduction to Pattern Recognition. Changsha: Changsha Institute of Technology Press, 1991. 30~154 (in Chinese).

[7] Floyd, R. W. Algorithm 97: shortest path. Communications of the ACM, 1962,35(5,6):345.

[8] Aho, A. V., Hopcroft, J. E., Ullman, J. D. The Design and Analysis of Computer Algorithms. New York: Addison-Wesley Publishing Company, 1974. 189~195.

附中文参考文献:

[1] 汪培庄,李洪兴.模糊系统理论与模糊计算机.北京:科学出版社,1996.166~191.

[2] 贺仲雄.模糊数学及其应用.天津:天津科学技术出版社,1984.76~185.

[3] 何新贵.模糊知识处理的理论与技术(第2版).北京:国防工业出版社,1998.414~421.

[4] 张跃.模糊数学方法及其应用.北京:煤炭工业出版社,1992.273~333.

[5] 陈文伟.智能决策技术.北京:电子工业出版社,1998.9~21.

[6] 沈清,汤霖.模式识别导论.长沙:国防科学技术大学出版社,1991.30~154.

An Optimal Algorithm for Fuzzy Classification Problem*

MA Jun[1], SHAO Lu[2]

[1](*Department of Computer Science, Shandong University, Ji'nan 250100, China*);

[2](*Shandong Medicine Industry Institute, Ji'nan 250100, China*)

E-mail: majun@cs.sdu.edu.cn

http://www.cs.sdu.edu.cn

Abstract: In this paper, the geometric meaning of the transitive closure of a fuzzy relation in corresponding fuzzy graph is first given. An optimal algorithm, which is based on the computation of graph connected components, for fuzzy classification problem is proposed. For any given n samples, the worst case time complexity $T(n)$ of the algorithm satisfies that $O(n)\leqslant T(n)\leqslant O(n^2)$. Compared with the classic fuzzy classification algorithm, which is based on the computation of the transitive closure of a given relative matrix and of the $O(n^3\log n)$ time, the new algorithm decreases $O(n\log n)$ time factor at least. The theoretic analysis and computer performance show that the real computing time of the new algorithm is acceptable when it is used for fuzzy classification on large data.

Key words: fuzzy theory; fuzzy relation; fuzzy classification; fuzzy application

* Received September 14, 1999; accepted December 8, 1999

Supported by the National High Technology Development Program of China under Grant No. 863-306-ZT06-01-4; the Natural Science Foundation of Shandong Province of China under Grant No. Z99G01