CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2012 年第 34 卷第 12 期
Vol. 34, No. 12, 2012

文章编号:1007-130X(2012)12-0155-05

基于模糊逻辑的 K-means 算法研究*

Research of K-means Algorithm by Fuzzy Logic

陈苏蓉,朱晓辉

CHEN Su-rong, ZHU Xiao-hui

(南通大学计算机科学与技术学院,江苏 南通 226019)

(College of Computer Science and Technology, Nantong University, Nantong 226019, China)

摘　要:K-means 算法的基本思想是通过迭代方法把所有的元素都唯一聚类到不同的簇中,使得同一簇中的质点具有最小相异度,不同簇间的元素具有最大相异度。但是,这种聚类方法使得那些属于不同簇的交叉区域中的质点也被简单地聚类到了某个簇中,因此无法表达某些元素的跨簇特性。本文提出了基于模糊逻辑的 K-means 算法,利用模糊逻辑来计算不同簇交叉区域中质点属于某个簇的权重,在获得聚类结果的同时可以有效描述质点的跨簇特性。实验结果表明该算法是有效的。

Abstract: The basic idea for the K-means algorithm is to partition all elements to different clusters by iterative method so that the elements in the same cluster have the minimum dissimilarity and the elements in different clusters have the maximum dissimilarity. However, it may simply cluster these elements in an overlap which should be in different clusters to the same cluster, so the result of clustering cannot show the element's overlap characteristic. In the paper, a new K-means Algorithm based by fuzzy logic is proposed. It can not only obtain the same clustering result as original algorithm but also get element's overlap characteristic. Experiment shows that the new algorithm is efficiency.

关键词:模糊逻辑;K-means 算法;跨簇特性

Key words: fuzzy logic; K-means algorithm; overlap characteristic

doi: 10.3969/j.issn.1007-130X.2012.12.027

中图分类号:TP302　　**文献标志码**:A

1　引言

聚类分析是数据挖掘的一项基本任务,目标是将数据聚集成不同的分类[1]。K-means 算法是一种典型的基于划分的数据挖掘聚类算法[2,3],其基本思想是给定一个元素集合 D,每个元素具有 m 个可观察属性,使用某种算法将 D 划分成 k 个子集,要求每个子集内部的元素之间相异度尽可能低,而不同子集的元素相异度尽可能高,其中每个子集叫做一个簇。一般认为 K-means 算法具有如下两点不足:(1)对初始质心的选择敏感,即针对同一个数据集中的元素,不同的输入次序、选择不同的初始质心会得到不同的聚类结果。(2)集合中的异常点会对聚类结果产生较大的影响。针对以上问题,很多文献结合各种智能优化算法(如蚁群算法、遗传算法等)对 K-means 算法进行改进并取得了较好的结果。但是,由于 K-means 算法的特性,使得那些属于不同簇的交叉区域中的质点也被简单地聚类到了某个簇中,从而失去了该质点的跨簇

* 收稿日期:2011-12-26;修订日期:2012-02-27
基金项目:南通市基础应用研究项目(K2010067,BK2011072);江苏省科技厅应用研究项目(BC2010122);江苏省高校科研成果产业化推进项目(JHB2011-45);江苏省"六大人才高峰"项目(2010-WLW-006);南通市重大科技创新专项计划项目(XA2008004);江苏高校优势学科建设工程资助项目
通讯地址:226019 江苏省南通市南通大学计算机科学与技术学院
Address: College of Computer Science and Technology, Nantong University, Nantong, Jiangsu 226019, P. R. China

特性。本文在 K-means 算法基础上提出了利用中心质点来消除异常数据对聚类结果的影响,并引入了模糊逻辑理论,利用模糊逻辑来计算不同簇交叉区域中的质点属于某个簇的权重。通过实验证明,该方法不但可以消除异常质点对聚类结果的影响,还可以很好地描述质点的跨簇特性。

2 K-means 算法及模糊逻辑理论

2.1 相异度计算

定义 1(相异度) 设元素集合为 $D,\{X_1,X_2,\cdots,X_i,\cdots,X_n\}$ 为 D 中的元素$(1\leqslant i\leqslant n)$,$n$ 为集合中元素的总数,每个元素具有 m 个可度量特征属性。设 X_i、X_j 是两个元素项,那么 X_i 和 X_j 的相异度为:$d(X_i,X_j)=f(X_i,X_j)\rightarrow \mathbf{R}$,其中 R 为实数域。

相异度是两个元素对实数域的一个映射,所映射的实数定量表示为两个元素的相异度。在 K-means 算法中常以欧几里德距离表示质点间的相异度,其定义为:

$$d(X_i,X_j)=\sqrt{(X_{i1}-X_{j1})^2+\cdots+(X_{im}-X_{jm})^2} \tag{1}$$

其中,X_i 和 X_j 表示集合 D 中第 i 和第 j 个元素,m 为元素的可度量特征数。

2.2 K-means 算法描述

K-means 算法的执行过程如下:

(1)从集合 D 中随机取 k 个元素,作为 k 个簇各自初始的质心,k 为事先设定的分类数。

(2)分别计算集合 D 所有元素到 k 个簇中质心的相异度 $d(X_i,X_j)$,将这些元素分别划归到相异度最低的簇中。

(3)根据聚类结果,重新计算所有簇中元素各自维度的算术平均数,作为各簇新的质心。

(4)转步骤(2),将 D 中全部元素按照新的质心重新聚类,直到聚类结果不再变化。

(5)将结果输出。

2.3 模糊逻辑理论

模糊逻辑也称为模糊集合理论,自 Zadeh 于 1965 年提出模糊集合理论以来,模糊理论得到了广泛应用[4]。它不同于传统的集合理论,在传统集合中用功能函数来描述集合,如果该元素属于集合则功能函数的值为 1,否则为 0。在模糊逻辑中用成员关系函数来定义模糊集。模糊逻辑的优点在于它能够描述同时隶属于不同分类范围的概念[5]。比如用来描述学生成绩的好和差两种状态分别用 1.0 和 0.0 来表示,这里没有一个非常明确的分界点来描述成绩的好和差,因此很难用传统的集合理论来描述,而用模糊集理论则可以很方便地描述。如图 1 所示,可以看到用模糊集理论可以允许学生的成绩介于好和差之间。

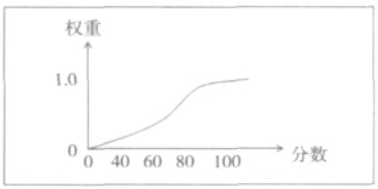

图 1 学生成绩描述

3 模糊集理论在 K-means 算法中的应用

3.1 构造成员关系函数

实际应用中,集合中的元素一般都有时间相关性,比如一个心脏病人在一段时间内的心率数据,其中某时刻的心率数据和该时刻前后的两段心率数据有一定的相关性。因此,该时刻心率数据属于某个簇的可能性跟该数据前后两段数据在某个簇中的可能性有很大的相关性。根据前面相异度的定义,现假定元素 X_i 位于两个簇的交叉区域,设 v 为元素 X_i 位置前后相邻元素的个数,则 $X_{i-v},\cdots,X_{i-1}$ 为 X_i 的前 v 个元素,$X_{i+1},\cdots,X_{i+v}$ 为 X_i 的后 v 个元素。通过对这 $2v$ 个元素分别赋予不同的权重值来表示对元素 X_i 的影响程度。假定该 $2v$ 个元素中分别有一部分落在两个簇 C_1 和 C_2 中,则可以通过分别统计在 C_1 和 C_2 簇中的元素对元素 X_i 的权重 W 来度量 X_i 在 C_1 和 C_2 簇中的权重。计算一个元素 X_t 对 X_i 的权重如公式(2)所示:

$$W_t=\frac{v+1-|t-i|}{v!} \tag{2}$$

其中,$i-v\leqslant t\leqslant i+v$,并且 $t\neq i$,i 表示元素 X_i 的编号,v 表示要统计的元素 X_i 前后相邻元素的数目,t 表示与 X_i 相邻近的元素编号。W_t 表示元素 X_t 对元素 X_i 的影响因子权重。从公式(2)可以看出,当 t 值越接近 i 时,W_t 的值就越大,说明元素 X_t 对元素 X_i 的影响就越大;相反,当 t 的值越远离 i 时,W_t 的值就越小,说明元素 X_t 对元素 X_i 的影响就越小。由此可以得出 X_i 在簇 C_k 中的权重如公式(3)所示:

$$W_iC_k=\sum_{l=1}^{L}W_l \tag{3}$$

其中,L 是与元素 X_i 相邻的 $2v$ 个元素中落在 C_k 簇中的所有的元素集合,l 是元素的编号。

3.2 确定 K-means 算法初始质心

由于集合中的异常元素会对 K-means 算法的聚类结果产生较大影响,因此需要先排除异常元素,然后才能进行聚类。这里采用先计算集合中所有元素的中心点;然后再计算所有元素到该中心点的距离并排序,把距离中心点最远的前几个元素作为异常点排除;最后进行 K-means 聚类。将中心质点记作 X_c,根据前面相异度的定义,X_c 的 m 维特征属性分别为 $X_{c1},X_{c2},\cdots,X_{ci},\cdots,X_{cm}$。其第 i 维特征属性的计算如下:

$$X_{ci}=\frac{X_{1i}+X_{2i}+\cdots+X_{ni}}{n} \tag{4}$$

其中,$1\leqslant i\leqslant m$,由欧几里德距离公式(1)可知,求元素 X_i 到中心点 X_c 的距离公式为:

$$d(X_i,X_c)=\sqrt{(X_{i1}-X_{c1})^2+\cdots+(X_{im}-X_{cm})^2} \tag{5}$$

3.3 算法实现

基于模糊逻辑的 K-means 算法流程描述如下所示:

(1)利用公式(4)计算初始质心 X_c;

(2)利用公式(5)计算 D 中所有元素到 X_c 的距离,并求所有元素到 X_c 的平均距离 L;

(3)若某个元素到初始质心 X_c 的距离大于所有元素到初始质心 X_c 的平均距离 L 的两倍,则认为该元素是异常数据,从数据集中剔除;

(4)用 K-means 算法求剩余元素的聚类;

(5)聚类结果中落于交叉区域的元素放入集合 P;

(6)依次取 P 中元素,并利用公式(2)和公式(3)计算与该元素相邻的 $2v$ 个元素的权重,得到该元素属于不同聚类的权重;

(7)重复步骤(6)直至 P 中元素全部计算完毕;

(8)结束。

其中,判断某个元素是否属于交叉区域的算法为:依次取集合中的元素 X_i,若与该元素相邻的前 v 个元素和后 v 个元素与 X_i 不完全同属于一个聚类,则 X_i 元素为交叉区域的元素。

4 算法验证

这里以心率失常患者的心率数据作为实验数据。为不失一般性,同时考虑到数据集的规模,选取患者一天的心率数据,并以 15 分钟为时间单位,以患者 15 分钟内的平均心率数据为一个有效元素值,因此共有 $60/15*24=96$ 个样本数据。考虑到元素的时间特性,这里以每分钟平均心率和该心率的时间点(小时)作为元素的两个属性。若第一个 15 分钟的平均心率是 60,则元素值为(60, 0.25),第二个 15 分钟的平均心率是 63,则值为(63, 0.5),依次类推,元素集合如表 1 所示。

表 1 心率样本数据

所有心率样本元素			
66,0.25	67,6.25	83,12.25	75,18.25
61,0.5	69,6.5	75,12.5	72,18.5
64,0.75	68,6.75	79,12.75	69,18.75
69,1	62,7	84,13	71,19
90,1.25	79,7.25	86,13.25	69,19.25
64,1.5	78,7.5	86,13.5	70,19.5
69,1.75	76,7.75	76,13.75	60,19.75
63,2	76,8	78,14	72,20
68,2.25	78,8.25	85,14.25	72,20.25
69,2.5	81,8.5	79,14.5	60,20.5
65,2.75	81,8.75	88,14.75	69,20.75
61,3	75,9	58,15	73,21
67,3.25	82,9.25	86,15.25	63,21.25
64,3.5	79,9.5	90,15.5	69,21.5
68,3.75	82,9.75	75,15.75	64,21.75
64,4	73,10	81,16	72,22
69,4.25	67,10.25	81,16.25	61,22.25
61,4.5	65,10.5	78,16.5	85,22.5
62,4.75	63,10.75	78,16.75	66,22.75
63,5	65,11	82,17	75,23
67,5.25	69,11.25	81,17.25	73,23.25
63,5.5	69,11.5	84,17.5	63,23.5
62,5.75	69,11.75	77,17.75	62,23.75
70,6	73,12	87,18	62,24

最后在集合中随机添加(0,0)、(5,5)、(10,10)、(33,134)四个元素作为数据集中的噪声数据。我们把心率聚类分成三种情况:心动过速、心动正常和心动过缓,即取 K 值为 3。传统的 K-means 算法聚类结果如图 2 所示。

从图 2 聚类结果看出,由于传统 K-means 算法没有考虑如何去除噪声数据,因此聚类结果受到噪声数据的严重干扰,把四个噪声数据聚类为一类,把正常数据聚类为另外两类。

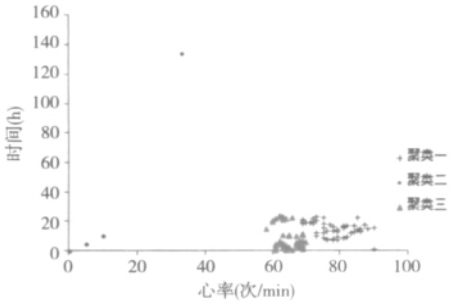

图 2 传统 K-means 算法聚类结果

用改进后的 K-means 算法进行聚类时,首先利用公式(4)求得数据集中心元素为(69.67,13.13)。再利用公式(5)计算各元素到中心元素的距离,然后计算得到平均距离为 $L=13.12$。最后分别判断每个元素到中心元素的距离是否大于平均距离的两倍,若是则把该元素从集合中删除。由公式(5)计算得到元素(0,0)、(5,5)、(10,10)、(33,134)到中心元素的距离分别为 70.89、65.17、59.75、126.31。这几个距离都大于平均距离 L 的两倍,因此算法自动把这四个元素作为异常元素从集合中排除。改进后的 K-means 算法的聚类结果如图 3 所示。

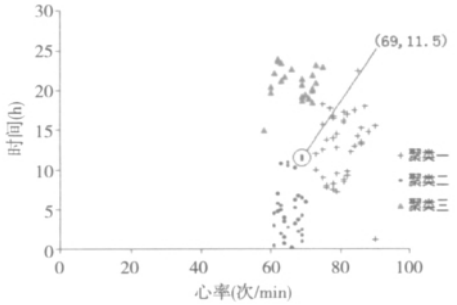

图 3 改进后 K-means 算法聚类结果

其中聚类一中元素如表 2 所示,聚类二中元素如表 3 所示,聚类三中元素如表 4 所示。

表 2 聚类一

聚类一元素集合			
90,1.25	79,9.5	76,13.75	78,16.5
79,7.25	82,9.75	78,14	78,16.75
78,7.5	73,10	85,14.25	82,17
76,7.75	73,12	79,14.5	81,17.25
76,8	83,12.25	88,14.75	84,17.5
78,8.25	75,12.5	86,15.25	77,17.75
81,8.5	79,12.75	90,15.5	87,18
81,8.75	84,13	75,15.75	75,18.25
75,9	86,13.25	81,16	85,22.5
82,9.25	86,13.5	81,16.25	

表 3 聚类二

聚类二元素集合			
60,0.25	65,2.75	63,5	67,10.25
61,0.5	61,3	67,5.25	65,10.5
64,0.75	67,3.25	63,5.5	63,10.75
69,1	64,3.5	62,5.75	65,11
64,1.5	68,3.75	70,6	69,11.25
69,1.75	64,4	67,6.25	69,11.5
63,2	69,4.25	69,6.5	69,11.75
68,2.25	61,4.5	68,6.75	
69,2.5	62,4.75	62,7	

表 4 聚类三

聚类三元素集合			
58,15	60,19.75	63,21.25	75,23
72,18.5	72,20	69,21.5	73,23.25
69,18.75	72,20.25	64,21.75	63,23.5
71,19	60,20.5	72,22	62,23.75
69,19.25	69,20.75	61,22.25	62,24
70,19.5	73,21	66,22.75	

设 $v=4$,由前面判断元素是否属于交叉区域的算法可知,圆圈中的点(69,11.5)处于聚类一和聚类二的交叉区域。由表 1 可知该点是集合中的第 46 个元素,考虑该点前后各 4 个元素的权重值,即分别计算 42、43、44、45 号元素以及 47、48、49、50 号元素对 46 号元素的权重。由表 2～表 4 可知,42、43、44、45、47 号元素在聚类二中,而 48、49、50 号元素在聚类一中。由公式(2)和公式(3)计算可知,该节点属于聚类一的权重:

$$W_iC_1=\frac{5+1-|48-46|}{5!}+\frac{5+1-|49-46|}{5!}+\frac{5+1-|50-46|}{5!}$$

该节点属于聚类二的权重:

$$W_iC_2=\frac{5+1-|42-46|}{5!}+\frac{5+1-|43-46|}{5!}+\frac{5+1-|44-46|}{5!}+\frac{5+1-|45-46|}{5!}+\frac{5+1-|47-46|}{5!}$$

最后计算得到该节点属于聚类一和聚类二的权重比为 9∶19。这也正好说明了该元素虽然属于聚类二,但也比较靠近聚类一的现状。

5 结束语

本文在分析了 K-means 算法不足的基础上提出了先求中心质点,然后利用各质点到中心质点的

距离的方法来去除异常质点,并利用模糊集理论来解决 K-means 算法中处于不同簇的交叉区域质点的跨簇特性问题。通过心脏病患者的数据进行实验证明,与传统的 K-means 算法相比,算法有效消除了噪声数据对聚类结果的影响,并能有效描述交叉区域中元素的跨簇特性。

参考文献:

[1] 张霞,王素贞,尹怡欣,等. 基于模糊力度计算的 K-means 文本聚类算法研究[J]. 计算机科学,2010,46(8):209-211.

[2] 王慧,申石磊. 一种改进的特征加权 K-means 聚类算法[J]. 微电子学与计算机,2010,27(7):161-163.

[3] 邱保志,郑智杰. 基于局部密度和动态生成网格聚类算法[J]. 计算机工程与设计, 2010,31(2):385-387.

[4] 王慧,张颖超. 基于模糊逻辑带权重的模糊查询研究[J]. 计算机应用研究,2009,26(1):214-216.

[5] 黄国言,常旭亮,高健培. 模糊逻辑理论在入侵检测系统中的应用研究[J]. 计算机工程与应用, 2010,27(5):110-112.

陈苏蓉(1977-),女,江苏宿迁人,硕士,讲师,研究方向为软件工程和分布式数据库。E-mail:ntuzxh@gmail.com

CHEN Su-rong, born in 1977, MS, lecturer, her research interests include software engineer and distributed database.

朱晓辉(1976-),男,江苏海门人,硕士,副教授,研究方向为软件工程和分布式系统。E-mail:Zhufirst@ntu.edu.cn

ZHU Xiao-hui, born in 1976, MS, associate professor, his research interests include software engineer and distributed system.

《计算机工程与科学》杂志投稿须知

1. 征稿范围

本刊刊登具有创新性、高水平、有重要意义的原始性研究学术论文以及反映学科最新发展状况的文献综述和信息性文章。来稿应观点明确,论据充分、数据可靠,层次分明,文理通顺。

2. 投稿要求和注意事项

(1) 文题、作者姓名(一般不超过 6 人)、作者单位及所在城市和邮编、摘要、关键词均需中英文对照。论文如果获得有关研究基金或课题资助 ,需提供基金名称及编号(亦需中英文对照)、并提供第一作者的姓名、性别、民族(汉族不写)、出生年、职称、学位以及联系人姓名、职称、电话、传真及 E-mail 地址。

(2) 论文题目应简洁、准确,不宜使用缩略词;摘要(中文)字数一般在 200~300 字间,内容应包括论文的研究目的、方法及研究结果等;英文摘要字数在 300~400 个单词左右(对中文摘要内容进行扩展),简要地介绍研究背景、研究内容、研究成果。关键词的个数为 3~8 个。正文字数不得低于 9000 字。

(3) 文中量、单位及符号的使用应符合国际标准和国家标准。注意容易混淆的外文字母的文种、大小写、正斜体及上下角标的正确书写。文中外国人名、术语统一为英文,不宜采用中文译法。

(4) 图、表和公式应通篇分别编号,图题、表题应有中英文对照。表格应采用三线表形式,内容以英文表述。

(5) 稿件具体格式:正文请按照五号宋体、通栏式排版。

3. 投稿约定

(1) 原稿必须是在中外文正式刊物上未发表的论文。本刊严禁一稿多投、重复内容多次投稿、不同文种重复投稿。一旦发现上述情况,稿件将按退稿处理,并将通知作者单位及有关期刊。作者本人的稿件今后将不被录用。

(2) 稿件审查结果在三个月内通知作者 ,在此其间 ,作者不得将稿件投往他处。个别稿件可能送审时间较长。如果作者决定改投他刊或退稿,请通知编辑部后,再进行处理。编辑部决定录用稿件后 ,将及时通知作者。

(3) 学术研究必须真实,投稿必须合法,即不存在抄袭、剽窃、侵权等不良行为。如发现上述不良行为,本刊将据实通知作者所在单位的最高领导层,并不再接受第一作者的投稿。作者文责自负,本刊不承担连带责任。

(4) 在稿件的修改过程中,若超过稿件修改时限 30 日,编辑部将以作者返回修改稿日期作为投稿日期;超过 30 日,编辑部有权对稿件做出退稿处理。

(5) 文责自负, 编辑部有权对稿件做技术性、文字性修改 , 在征得作者同意后可以进行实质内容的修改。

(6) 论文发表后 , 版权即属于编辑部所有(包括上网的版权)。

(7) 作者需交纳审稿费和发表费, 编辑部将给予一定的稿酬,同时赠寄当期杂志 2 册。