文章编号:1671-6833(2010)01-0089-04

基于加权欧式距离的k_means算法研究

张忠林,曹志宇,李元韬

(兰州交通大学 电子与信息工程学院,甘肃 兰州 730070)

摘 要:传统的k_means算法将欧式距离作为最常用的距离度量方法.针对基于欧式距离计算样本点与类间相似度的不足,用"相对距离"代替"绝对距离"可以更好地反映样本的实际分布,提出一种在领域知识未知的情况下基于加权欧式距离的k_means算法.针对公共数据库UCI里的数据实验表明改进后的算法能产生质量较高的聚类结果.

关键词:k_means算法;聚类;加权;变异系数

中图分类号:TP391 **文献标识码**:A

0 引言

数据库和信息技术的研究与发展以及获取数据手段的多样化,人们所拥有的数据量急剧增加,数据挖掘[1]引起了信息产业界的极大关注,其主要原因是存在大量数据,可以广泛使用,并且迫切需要将这些数据转换成有用的信息和知识.数据挖掘是从大型数据库或数据仓库中提取出可信、新颖、有效,并能被人理解模式的高级处理过程,可广泛应用于决策支持、信息管理、科学研究等领域.数据挖掘的任务包括分类预测、关联规则分析、聚类分析、时间序列模式和偏差分析等.其中聚类分析是研究与应用的热点之一,也是数据挖掘领域的一个重要分支.聚类就是一个将数据集划分为若干簇或类的过程,通过聚类使得同一类内的数据对象具有较高的相似度,而不同类中的数据对象具有较高的相异度.

聚类算法[2]大体可以划分为如下几类:划分的方法[3]、层次的方法、基于密度的方法、基于网格的方法、基于模型的方法、基于遗传算法的方法、基于蚁群算法的方法等.参考文献[4]对以上各种聚类算法的性能与适用范围做了分析.

1 k_means算法

1.1 k_means算法的思想

k_means算法以k为参数,把n个对象分成k个簇,以使簇内具有较高的相似度,而簇间的相似度较低.相似度的计算根据一个簇中对象的平均值来进行.相似度的定义是划分的关键.

k_means[5]算法的基本思想是:随机地选择k个对象,每个对象初始地代表了一个簇的平均值或中心.对剩余的每个对象,根据其与各个簇中心的距离,将它赋给最近的簇.然后重新计算每个簇的平均值.这个过程不断重复,直到目标函数收敛.通常定义为公式(1)的目标函数,采用启发式方法使得目标函数值最小.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i | \qquad (1)$$

式中:E为集合U中所有对象与相应聚类中心的均方差之和;p为对象空间中一个数据对象;m_i为类C_i的均值(p和m_i都是多维的).该函数旨在使生成的聚类结果簇尽可能地紧凑和独立.

由于该方法在聚类过程中采取距离就近原则,在实际应用中,不考虑数据样本中的每个属性变量在聚类过程中的不同作用,而是将它们统一看待.用样本之间的欧氏距离并不能准确地表示相似度,因为相似不仅仅依赖于样本间的相近程度,还依赖于数据样本间的内在性质,也就是说数据集中每个属性在聚类分析过程中对于数据样本划分的重要性不同.

1.2 k_means算法的一般过程

k_means算法属于划分方法,它需要有先验

收稿日期:2009-09-06;修订日期:2009-11-15

基金项目:兰州市企业技术攻关计划资助(2009-1-4);兰州交通大学"青蓝"人才工程基金资助(QL-05-10A)

作者简介:张忠林(1965-),男,河北阜城人,兰州交通大学教授,博士,主要研究方向:智能信息处理、软件工程,E-mail:zhangzl@mail.lzjtu.cn;曹志宇(1983-),男,内蒙古包头人,硕士研究生,主要研究方向:数据挖掘.

知识 k,即知道数据对象集合要聚为几类. 在此,笔者给出 k_means 算法的一般过程.

设两个 p 维向量 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 $x_j = (x_{j1}, x_{j2}, \cdots, x_{jp})$ 分别表示两个对象,它们的欧氏距离为:

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \tag{2}$$

下文介绍 k_means 算法的一般过程.

算法输入:聚类个数 k,以及包含 n 个数据对象的数据样本集 U;

算法输出:满足方差最小标准的 k 个聚类.

算法步骤:

(1) 从 n 个数据对象中任意选择 k 个对象作为初始聚类中心;

(2) 根据每个聚类中所有对象的均值(中心对象),计算样本集中每个对象与这些中心对象的欧式距离,并根据最小距离重新对相应对象进行划分;

(3) 重新计算每个(有变化) 聚类的均值(中心对象);

(4) 循环执行(2) 到(3),直到每个聚类不再发生变化为止.

2 基于加权欧式距离的 k_means 算法

2.1 变异系数赋权法

数据样本集用可分性较好的数据样本来描述,具有相同类别的数据样本越集中,而不同类别的数据样本越远离,表现在散点图上就是数据点的分散性比较好,而且类与类之间的距离比较大. 为了反映数据的离散程度,通过对多种赋权法的比较,选用变异系数作为其权值. 定义如下:[5]

变异系数赋权法是在方差倒数赋权法的基础上提出的. 一组数据的变异系数是它的标准差除以均值的绝对值. 即对数据集中的 n 个数据 x_1, $x_2, \cdots\cdots, x_n$. 记为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

$$S_x = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2\right)^{\frac{1}{2}} \tag{4}$$

则

$$v_x = S_x / |\bar{x}| \tag{5}$$

就是 $x_1, x_2, \cdots\cdots, x_n$ 的变异系数.

于是,对数据库中选用的属性 $z_1, z_2 \cdots\cdots, z_j$,利用被评价对象的数据,各个属性都有各自的变异系数. 为了方便,用 v_i 表示 z_i 的变异系数,$i = 1, 2, \cdots, p$,此时,属性 z_i 相应的权重系数 w_i:

$$w_i = v_i / \sum_{i=1}^{p} v_i,\ i = 1, 2, \cdots, p \tag{6}$$

v_i 的值大表示 x_i 在不同的对象身上变化大,区别对象能力强,所以应给予重视.

2.2 加权欧式距离的 k_means 算法

基于加权欧式距离的 k_means 算法的基本思想是在给出待测数据库以后,首先计算数据集中各个属性的权值,再计算数据样本之间的相似度时使用加权欧式距离,即:

$$d(x_i, x_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_p(x_{ip} - x_{jp})^2} \tag{7}$$

下文为改进算法的一般过程.

算法输入:聚类个数 k,以及包含 n 个数据对象的数据样本集 U.

算法输出:满足方差最小标准的 k 个聚类.

算法步骤:

(1) 从 n 个数据对象中任意选择 k 个对象作为初始聚类中心;

(2) 根据每个聚类中所有对象的均值(中心对象),计算样本集中每个对象与这些中心对象的加权欧式距离,并根据最小的加权欧式距离重新对相应对象进行划分;

(3) 重新计算每个(有变化) 聚类的均值(中心对象);

(4) 循环执行(2) 到(3),直到每个聚类不再发生变化为止.

3 与传统的 k_means 算法的比较

基于加权欧式距离的 k_means 算法与传统的 k_means 算法相比较,就是把传统算法中计算样本点与类间相似度的欧氏距离变成了这里的加权欧氏距离,计算加权欧式距离中的权值会花费一定的时间,但是它对数据集中样本的处理能力却大大提高了. 在处理数据的实际过程中,对"噪声"和孤立点样本不是十分敏感,少量该样本不会对权重系数的选取产生大的影响. 该算法的复杂度和传统算法是一致的,也是 $O(nkt)$,其中:n 是数据集中所有样本的数目;k 是簇的数目;t 是迭代的次数. 因此当 $k \ll n$ 且 $t \ll n$ 时,对处理大数据集是可伸缩的和高效率的. 另外,该算法和传统算法一样都需要事先估计簇的个数 k,如想得到最优解时必须试探不同的 k 值.

4　实验分析

为了验证改进的效果，笔者对传统的 k_means 算法、基于加权欧式距离的 k_means 算法进行了对比实验. 实验的平台是 Windows XP，P4 2.8 GHz CPU，512 MB 内存，80 GB 硬盘. 开发工具采用 VC + + 6.0. 选用 UCI 数据库上的 Iris、Wine、Balance - scale、New - thyroid、Haberman 五组数据集作为测试数据. UCI 数据库是一个专门用于测试机器学习、数据挖掘算法的公共数据库，库中的数据都有确定的分类，因此可以用准确率直观地表示聚类的质量. 整个实验过程中，对于传统的 k_means 聚类算法，选取不同的初始聚类中心，运行 10 次，和基于加权欧式距离的 k_means 算法进行比较. 如表 1 所示.

表 1　传统 k_means 算法与基于加权欧式距离的 k_means 算法实验结果

Tab. 1　The resules between the traditional k_means algorithm and the euclid distance with weights of k_means algorithm

数据集	样本个数	属性个数	聚类个数	k_means 准确率/%	加权 k_means 准确率/%
Iris	150	4	3	70.08	83.67
Wine	178	13	3	62.47	78.52
Balance-scale	625	4	3	48.29	50.96
New-thyroid	215	5	3	71.86	84.19
Haberman	306	4	2	53.86	56.77

Iris 数据集包含 150 条样本记录，分别取自三种不同的鸢尾属植物 Setosa、Versicolor 和 Virginica 的花朵样本，每一类各 50 条记录，其中每条记录有 4 个属性：萼片长度（Sepal length）、萼片宽度（Sepal width）、花瓣长度（Petal length）和花瓣宽度（Petal width）. 它们的取值范围分别为：(4.3 ~7.9)，(2.0 ~ 4.4)，(1.0 ~ 6.9)，(0.1 ~ 2.5). 所对应的权值分别为：0.126、0.125、0.196、0.552. 可以看出，同一个属性中的样本取值范围较为接近时，这个属性所对应的权值也较小；相反，同一个属性中的样本取值范围较大时，对应的权值也较大. 在利用改进后的 k_means 算法来计算样本相似度时，加权欧式距离可以比较准确地代表它们的数据分布，所以对数据集的划分也较为准确，最终得到了很好的聚类结果.

Wine 数据集中的数据为产于意大利同一地区不同种植园的 3 种葡萄酒的成分分析样本，包含 178 条样本记录，分成 3 类，每个数据项有 13 个属性，各个属性的取值范围差距较大，可由这些属性确定葡萄酒出产的种植园. 经过分析发现：对分类最为重要的包括 3 个属性：Alcohol，Flavanoids，Color_intensity，它们的取值范围分别为：(0.34 ~ 5.08)，(1.28 ~ 13)，(11.03 ~ 14.83)，而对分类贡献不大的属性 Proline 的取值范围是 (278 ~ 1 680)，在计算距离时，属性 Proline 起到了绝对的作用，从而导致用欧氏距离来计算样本间的相似度并不能很好地代表数据的实际分布情况. 所以传统的 k_means 算法在对 Wine 数据集进行处理的过程中，受属性 Proline 影响，样本间相似度的计算产生了巨大的偏差，最终的聚类结果不令人满意，仅有 62.47%. 在引入加权欧式距离来计算相似度后，可以减少属性 Proline 对数据样本的影响.

传统的 k_means 算法没有考虑到数据的实际分布情况，而只是给出了一个算法可以运行的必要条件（把样本间的欧式距离作为相似度）. 应用改进后的算法，由于在一开始就计算得到了各个属性的权值，并将权值引入到计算欧式距离的过程中，所以得到了一个较为准确的样本间的相似度，在所选的五组数据集中，加权后有三组数据的准确率较传统算法的准确率获得明显上升，Balance - scale、Haberman 数据的准确率也略微有所提高.

5　结论

传统欧氏距离的计算方法，认为数据集中的所有属性在聚类中作用是相同的，用这种方法计算的欧式距离不能够准确反映样本之间的相似度基于这种情况，笔者提出的基于变异系数加权欧氏距离的计算方法，通过加权充分体现各个属性在聚类中的重要性，从而提高聚类结果的准确性和有效性. 实验结果表明，变异系数加权欧氏距离的方法克服了传统欧氏距离在聚类算法中的缺陷，优化了算法的性能.

参考文献：

[1] PANG N T, MICHAEL S, VIPIN K. 数据挖掘导论（英文版）[M]. 北京：人民邮电出版社，2006.

[2] HAN J W, MICHELINE K. 数据挖掘概念与技术 [M]. 北京：机械工业出版社，2005.

[3] 沈洁，赵雷，杨季文，等. 一种基于划分的层次的聚类算法 [J]. 计算机工程与应用，2007，43(31)：175 - 177.

[4] 孙吉贵，刘杰，赵连宇. 聚类算法研究 [J]. 软件学

报,2008,19(01):48-61.
[5] 马卫武,李念平,杨志昂. 室内空气品质综合评价权重系数的确定与分析[J]. 通风除尘,2004,(11):9-11.

Research Based on Euclid Distance with Weights of K_means Algorithm

ZHANG Zhong-lin, CAO Zhi-yu, LI Yuan-tao

(School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: Euclid distance is commonly used to measure distance in the traditional k_means algorithm. The k_means algorithm based on weighted Euclid distance is researched and presented to overcome the existing problems of similarity calculation in clustering analysis based on traditional Euclid distance when we have no any domain knowledge about the data objects, the relative distance but not absolute distance is more accurately response to data distribution. Experiments on the standard database UCI show that the proposed method can produce a high accuracy clustering result.

Key words: k_means algorithm; clustering; weight; coefficient of variation

---

12月6日,2009年中国机器人大赛暨RoboCup中国公开赛第二分区赛在湖南长沙圆满完成各项比赛,北京大学、浙江大学等全国54所高校,167支代表队,360多名机器人技术选手参赛,通过参加4大类(机器人水球比赛、机器人武术擂台比赛、微软足球仿真机器人比赛及RoboCup救援机器人仿真比赛)15个项目的激烈角逐,决出了各个项目的冠、亚、季军。我校信息工程学院学生组队参加,取得了优异成绩,参赛队信工一队、信工二队包揽了机器人武术擂台赛类人组冠亚军,另外2支队伍分别获得微软足球机器人仿真3D类人机器人仿真冠军和轮式机器人仿真亚军。我校学工部、研工部对此项赛事非常关注,有关负责人亲自到场观摩指导。

另据悉,第十届"广茂达杯"中国智能机器人大赛于11月14日-15日在上海工程技术大学举行,大赛来自全国24个省、直辖市的350多支代表队,共一千二百多名青少年和机器人爱好者参加,参赛队伍与人数均为历史之最。教育部、科技部、公安部及三十多所大学机器人领域内的领导、顶级学者出席了大赛开幕式。我校信息工程学院组队参加,并获得了大赛二等奖。