基于 k – means 聚类算法的研究

黄 韬,刘胜辉,谭艳娜
(哈尔滨理工大学 计算机科学与技术学院,黑龙江 哈尔滨 150080)

摘 要:分析研究聚类分析方法,对多种聚类分析算法进行分析比较,讨论各自的优点和不足,同时针对原 k – means 算法的聚类结果受随机选取初始聚类中心的影响较大的缺点,提出一种改进算法。通过将对数据集的多次采样,选取最终较优的初始聚类中心,使得改进后的算法受初始聚类中心选择的影响度大大降低;同时,在选取初始聚类中心后,对初值进行数据标准化处理,使聚类效果进一步提高。通过 UCI 数据集上的数据对新算法 Hk – means 进行检测,结果显示 Hk – means 算法比原始的 k – means 算法在聚类效果上有显著的提高,并对相关领域有借鉴意义。

关键词:数据挖掘;聚类算法;k – means 算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673 – 629X(2011)07 – 0054 – 04

Research of Clustering Algorithm Based on K – means

HUANG Tao,LIU Sheng – hui,TAN Yan – na
(Sch. of Computer Sci. and Tech.,Harbin Univ. of Sci. and Tech.,Harbin 150080,China)

Abstract:Analyze and research the method of cluster analysis,analyze and compare many kinds of algorithms of cluster analysis,discuss their respective strengths and weaknesses. At the same time,according to the weaknesses of the cluster result of original k – means algorithm is significant influence by selecting the initial cluster centers randomly,a modified algorithm is proposed. Through taking sample many times to data set,choose final superior cluster center,bring down the impact of initial cluster centers to improved algorithm greatly. Simultaneously,the initial data is standadized once the initial cluster center is selected,makes cluster effect improved furthermore. Detecting new algorithm Hk – means through the date of UCI data set,the result shows that Hk – means algorithm is more prominent improved than initial k – means algorithm in cluster effect,and it's useful for conference to relative field.

Key words:data mining;clustering algorithm;k – means algorithm

0 引 言

数据挖掘(Data Mining)[1]是一种数据分析处理技术,一般采取排出专家因素而通过自动的方式来发现数据仓库、大型数据库或其他大量信息中新的、不可预见的或隐藏的有价值的知识。数据挖掘方法既可以是数学的,又可以是非数学的;既可以是归纳的,又可以是演绎的。发现的知识能被用于信息管理,查询优化,决策支持以及过程控制等,用在数据自身的维护,所以数据挖掘技术是当前多个领域研究的热点。

聚类分析[2]是数据挖掘领域中的一个重要分支,它既可以作为数据挖掘中其他分析算法的一个预处理步骤,也可以作为一个单独的工具发现数据库中数据分布的一些深入的信息。聚类算法大致分为划分方法,层次方法,基于密度方法,基于网格方法和基于模型方法。一般情况下,为了使聚类效果更佳,经常将划分方法作为其他聚类方法的预处理步骤,所以划分方法的好坏直接影响结合聚类算法的效果。k – means 算法是划分方法中应用最广泛的一种方案,所以改进 k – means 算法,不但改善了划分方法本身的性能,还对结合的聚类方法提供了良好的接口。

由于 k – means 算法选取初始聚类质心是随机的,导致聚类结果不稳定。为了提高聚类结果的稳定性,在传统的聚类方法上结合聚类融合思想,提出一种新的 Hk – means 聚类方法,进行多次采样,选取最终较优的初始聚类中心,使得改进后的算法受初始聚类中心选择的影响度大大降低,提高聚类结果的稳定性。

1 相关工作

聚类算法是一种非监督机器学习算法,其实质就是对人们事先不了解的数据集进行分组,使得同一组内的数据尽可能相似而不同组内的数据尽可能不同,其目的是揭示数据分布的真实情况。聚类分析在统计

收稿日期:2010 – 12 – 01;修回日期:2011 – 03 – 02
基金项目:哈尔滨市后备带头人基金项目(2004AFXXJ039)
作者简介:黄 韬(1982 –),男,黑龙江人,硕士研究生,研究方向为企业智能计算;刘胜辉,教授,硕士研究生导师,研究方向为计算机集成制造系统,企业智能计算。

数据分析、模式识别、图像处理、生物学以及市场营销等领域也有着广泛的应用前景。目前的聚类算法大致上主要被分为五类[3]: 划分方法,层次方法,基于密度的方法,基于网格的方法和基于模型的方法。

1.1　划分聚类算法

划分聚类的算法过程[4]如下: 给出一个 n 个对象或元组的数据库,一个划分方法构建数据的 k 个划分,每个划分即表示一个聚簇,并且 $k < n$。也就是说,它将数据划分为 k 个组,同时满足以下的要求: (1) 每个组至少包括一个对象; (2) 每一个对象必须属于且仅属于一个组。其思想是给定一个 n 个对象的数据库,通过迭代重定位策略优化特定的目标函数,尝试确定数据集的 k 个划分,每个划分表示一个聚簇($k \leqslant n$)。要求簇间的对象尽可能相近或相关,不同的簇间的对象尽可能不同。划分聚类算法实质上通过迭代重定位策略优化特定的目标函数,尝试确定数据集的一个划分[5,6]。它具有挖掘算法简单、速度快等优点,适用于中小规模的数据集中发现球状簇,但不能发现形状任意和大小差别很大的类簇,只能保证收敛到局部最优,聚类结果受初始聚类质心的影响,且对噪音点敏感。典型的代表算法是 k – means 算法。

1.2　层次聚类算法

层次方法对给出的数据集合而进行层次分级,叫做树聚类算法。它使用数据的联接规则,透过一种层次架构方式,反复将数据进行分裂或聚合,以形成一个层次序列的聚类的问题解。根据分解而形成的过程,层次聚类则可以分为凝聚的和分裂的[7]。凝聚的方法,也叫做自底向上方法,首先将每一个对象作为单独的一个组,接着相继地合并相近的对象和组,直到所有的组合都合并为一个,或者达到了一个中止条件。分裂的方法,也叫做自顶向下的方法,首先将所有的对象都置于一个簇中。在迭代的每一个步骤中,一个簇则被分裂为更小的簇,一直到最终每一个对象在单独的一个簇中,或达到了一个终止条件。一种纯粹的层次聚类方法的缺点主要在于一旦合并或者是分裂执行,则将不能修正,也就是说,如果某个合并或是分裂效果在后来证明是不好的选择,该方法无法退回或是更正。层次聚合算法的计算复杂性为 $O(n^2)$,适合于小型数据集的分类。

1.3　基于密度的聚类算法

基于密度的聚类方法是将数据的对象之间的距离与某一给定范围内数据对象的个数这两个参数相结合,得出“密度”的概念,然后按照密度的稠密程度来判定数据对象的聚集情况。即将簇看作是数据空间中被低密度区域分割开的稠密数据区域。它能从含有噪声的空间数据库中发现任意形状的聚类,但需要直接面对整个样本集进行操作,测试每个对象是否是核心对象,并对每个核心对象搜索其直接密度可达的对象。而由于密度连通关系的传递性[8],往往使得绝大多数的样本点聚集到非常少的几个类簇中(通常是一类)。在没有空间索引辅助下,算法复杂度为 $O(n^2)$。

1.4　基于网格的聚类算法

基于网格的聚类算法使用了一种多分辨率的网格的数据结构,把对象的空间量化为有限数目的单元,进而形成了一个网格结构。这种方法的优点是处理速度很快。但是在构建父亲单元的时候则往往没有考虑到子女单元及其相邻的单元之间的关系,从而降低了聚类结果的质量及其精确性。

1.5　基于模型的聚类算法

基于模型的聚类算法为每个簇假定一个模型,寻找数据对象与给定模型之间的最佳拟合。基于模型的算法可能通过构建反映数据的对象空间分布的密度函数来定位聚类,或者基于标准的统计数字来自动决定聚类的数目,从而产生健壮的聚类方法。典型的代表有 EM 算法,概念聚类以及神经网络。EM 算法比较易于实现并且算法简单,但是不能达到全局最优; 概念聚类在机器学习中有着广泛的应用,但是对于大型数据集聚类没有良好的伸缩性。

2　Hk – means 聚类算法

k – means 算法是划分聚类算法的典型代表之一,它具有算法简单,速度快等优点,经常作为其他算法的预处理步骤,所以它的精度直接或间接地影响着聚类结果的质量。由于原始的 k – means 算法对初始质心的选取敏感,导致结果质量不稳定。下文将针对这个问题提出解决办法。

2.1　k – means 算法

k – means 算法是划分聚类算法的典型代表,实质上该算法基于簇中对象的平均值。为了能达到全局最优,基于划分的聚类要求穷举所有可能的划分。算法的处理过程如下:

算法的输入: 数据库中的对象与簇的数目 k。

算法的输出: 使得平方误差准则最小的 k 个簇。

方法如下:

(1) 从整个样本 n 中,任意选择 k 个对象作为初始的簇的中心 m_i($i = 1,2,\cdots,k$)。

(2) 利用公式 1,计算数据集中的每个 p 到 k 个簇中心的距离 $d(p, m_i)$。

(3) 找到每个对象 p 的最小的 $d(p,m_i)$,将 p 归入到与 m_i 相同的簇中。

(4) 遍历完所有对象之后,利用公式 2 重新计算 m_i 的值,作为新的簇中心。

(5) 重新将整个数据集中的对象赋给最类似的簇。这个过程反复进行直至平方误差准则最小。

公式 1: $d(i,j) = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{in}-x_{jn})^2}$

其中 $i=(x_{i1},x_{i2},\cdots,x_{in})$ 和 $j=(j_{i1},j_{i2},\cdots,j_{in})$ 是两个 n 维数据对象。

公式 2: $m_k = \sum_{i=1}^{N} x_i/N$

其中的 m_k 代表第 k 个簇的簇中心,N 代表第 k 个簇中数据对象的个数。

平方误差准则试图使聚类结果尽可能地独立和紧凑,即簇内对象的相似度尽可能的高。定义如下:

$$E = \sum_{i=1}^{k}\sum_{p\in C_i} |p-m_i|^2$$

其中 E 表示所有对象的平方误差的总和,p 代表空间中的对象,m_i 代表簇 C_i 的平均值。

2.2 改进的 k – means 算法

算法思想: 为了降低初值对聚类结果的影响,提高 k – means 聚类结果的稳定性,文中将对样本集进行 h 次随机采样,再对各采样的样本集进行以 $k'(k'\geqslant k)$ 个质心的 k – means 运算,即得到 h 组在随机样本上的聚类结果,每组聚类结果包括 k' 个类簇中心,共 $h*k'$ 个簇。

文献 [8] 给出,如果采样过程是以非常随机的方式进行的,认为这些选取出来的样本足以代表整个样本集。所以通过分析采样样本集的数据分布情况就可以找到整个样本集全局较优的 k 个质心,从而使得聚类结果的稳定性提高。其中参数 h 为采样次数,实验表明 $h=3\sim10$ 通常已经足够。参数 k' 为各采样的样本集上质心数目,$k'>>k$。原始的 k – means 算法,在 k 的选取上可以被认为是随机爬山算法,对于具有大量局部极高点和 k 个全局最高点的数据集,k – means 算法很可能陷入局部极高点。然而类别数目 k 的增大类似于开辟更多的爬山路径,如果 $k'>>k$ 个路径,则到达全局最高点的可能性自然会增大。但同时随着 k' 的增大,计算的代价也会增大。在文献 [9] 中研究了 k 近邻分类算法中参数 k 的设置,给出一个法则 $k=n^{3/8}$。文献 [10 ~ 12] 推荐 k' 值的选取满足如下约束: $k'=\min(a\lfloor n/|u_{\min}|\rfloor, n^{5/8})$,其中,$u_{\min}$ 是需要关注的最小类簇,a 值一般小于等于 10。需要说明的是,在满足上述约束条件时,k' 的取值对于聚类结果不敏感。处理流程如图 1 所示。

然后再对这 $h\times k'$ 个聚类中心采取以下方法: 先找到密度最大的质心,将其放入集合 S 中,再计算其他质心与 S 集合中所有对象距离之和最大的质心,也将其放入到 S 中,直到选出 k 个聚类中心为止即集合 S 中有 k 个元素。再以这 k 个质心作为 k – means 原始中心,进行 k – means 运算,处理流程如图 2 所示。

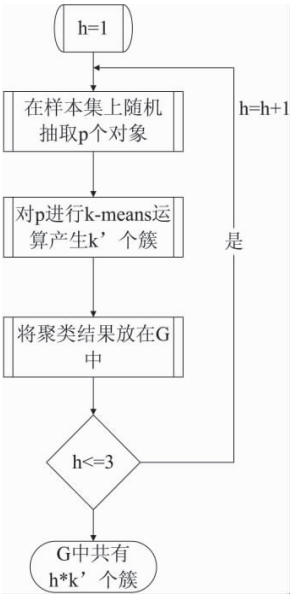

图 1 输出 $h*\ k'$ 个簇流程图

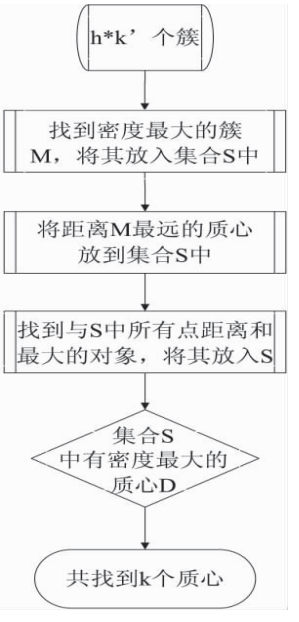

图 2 输出较优 k 个初始质心流程图

2.3 算法的复杂度分析

Hk – means 算法中,h 次预聚类时间复杂度为 $O(h\,tk'dn_{sum})$,其中,n_{sum} 为随机样本的大小,t 为迭代次数,d 为特征属性数,通常 k',$t<<n_{sum}$。再计算 $h\,k'$ 的密度,找到最大的簇,及最优的 k 个质心,时间复杂度为 $O((h\,k')^2)$,则 Hk – means 的时间复杂度为 $O(t(hk')^2dn)$,其中 n 为数据集的大小。大型数据集聚类时,Hk – means 算法比层次聚类算法快得多。

3 Hk – means 算法的应用

3.1 Hk – means 聚类算法的流程

Hk – means 算法的描述如下:

输入: 数据集 D、数据分类数目 k,采样次数 h;

输出: k 个簇。

方法:

(1) For $h = 1$ 对数据集 D 进行随机的采样,得到采样数据集 D';

(2) 利用 FindCenter(D',k'),获取 k' 个初始聚类中心{C_1,C_2,…,C_3};

(3) 将 k' 个样本点值分别赋给初始聚类中心 m_i,$1 \leqslant i \leqslant k'$;

(4) 利用公式 1,计算数据集 D' 中的所有点 $P_j(j = 1,2,\cdots,n_{sum})$,到 k' 个簇中心的距离 $d(p, m_i)$;

(5) 找到对象 p 的最小的 $d(p, m_i)$,将 p 归入到与 m_i 相同的簇中;

(6) 遍历完数据集 D' 中的所有对象之后,利用公式 2 重新计算 m_i 的值,作为新的簇中心;

(7) 重新将整个数据集中的对象赋给最类似的簇。这个过程反复进行直至平方误差准则最小,将 k' 个簇进行标记存储;

(8) 再一次重新采样,反复执行步骤 2 – 7 的过程。直到 h 与用户输入的参数相符。

(9) 利用 FindDesity(h C_k) 找到 k' 簇中密度最大的簇 C_i,将其放入集合 S 中,作为用户定义的最初始 k 个簇类中心的第一个成员;

(10) 利用 DistanceMean(S),得到集合 S 中的各簇中心的均值 m_i,再利用 Distance(m_i,m_j),找到不在集合 S 中其他簇类中心到 m_i 的最大值,将其归入集合 S 中;

(11) 反复计算 9 – 10,直到 S 中的元素个数为 k;

(12) 将集合 S 中的 k 个质心作为聚类的初始质心,对整个数据集进行步骤 4 – 7;

(13) 输出满足均方误差函数值最小的 k 个簇。

3.2　算法在审计中的应用

本节通过实验对 Hk – means 算法的有效性,及聚类质量进行分析。测试数据集,选取某省某市的单位实缴信息表,费款属期在 2007 年 1 月至 2007 年 6 月,总计 6 个月的数据为实验数据,数据条数共计 33096 条,对非数值属性值做了预处理,TestDB 为合成实验数据集。算法采用 JAVA 编写,在 Pentium(R) 4. 3. 00GHz,1GB 内存,160GB 硬盘,JBulider9. 0 环境下运行。

数据挖掘算法几乎都包含或多或少的参数,这些预先给定的参数值在很大程度上决定了数据挖掘的结果。如果参数值不符合数据的分布特征,就很难获得好的聚类结果。而在实际应用中,合适的参数值的确不好确定,一般采用专家知识或进行多次试验的方法,来取得一个参数的最佳近似值。

当 k 分别取 [4, 8,15,20,25] 时,实验结果见表 1。

表 1　k 值取不同值的结果

效果 算法 / 准确率 / k 取值	Hk – means	k – means
4	93	81
8	94	86
15	99	99
20	95	81
25	92	83

由表 1 可见 Hk – means 与 k – means 算法在 k 值等于 15 的情况下,可准确识别出绝大部分数据的类别。在其他 k 值情况下会出现一定差别,通过跟踪差别数据对象,发现造成差别主要原因是: k – means 算法随意选择初始的聚类中心,使得聚类效果时好时坏,而 Hk – means 算法对初始质心的选择进行了分析、计算,使得一般下聚类效果都好于原始的聚类算法。

4　结束语

文中通过将将聚类融合思想与 k – means 算法有机地结合,提出一种 Hk – means 算法,对数据集进行多次采样,选取最终较优的初始聚类中心,使得改进后的算法受初始聚类中心选择的影响度大大降低; 同时,在选取初始聚类中心后,对初值进行数据标准化处理,使聚类效果进一步提高。理论分析和实验结果表明,算法是有效可行的。下一步,将对减小参数点的选取对聚类质量的影响以及算法在高维空间数据集的应用做进一步研究。

参考文献:

[1] Tan Pang – Ning,Steinbach M,Kuma V. Introduction to Data Mining [M]. 北京: 人民邮电出版社,2006: 5 – 28.

[2] HAN Jiawei. Data Mining Concepts and Techniques [M]. 北京: 机械工业出版社,2006: 102 – 129.

[3] 孙吉贵,刘　杰. 聚类算法研究 [J]. 软件学报,2008,19(1): 48 – 61.

[4] 雷小峰,谢昆青. 一种基于 k – means 局部最优性的高效聚类算法 [J]. 软件学报,2008,19(7): 1683 – 1692.

[5] 周水庚,周傲英. 一种基于密度的快速聚类算法 [J]. 计算机研究与发展,2000,37(11): 1287 – 1292.

[6] 聂跃光,陈立潮,陈　湖. 基于密度的空间聚类算法研究 [J]. 计算机技术与发展,2008,18(8): 91 – 94.

[7] 赵　伟,张　姝,李文辉. 改进 K – means 的空间聚类算法 [J]. 计算机应用研究,2008,25(7): 1995 – 1997.

[8] 毕　华,梁洪力,王　珏. 重采样方法与机器学习 [J]. 计算

(下转第 62 页)

基于数据包标记技术的 FBT 和 PNM 溯源操作的收敛首先需要基站收集到足够的数据包，然后按相应的方法对数据包进行处理和计算得到完整的转发路径和数据源节点。其中，FBT 双层标签的使用可以帮助使用者不需要收集很多的数据就能找出攻击者所在分簇的头节点并快速构造出攻击的主干路径；基于日志记录技术的 CAPTRA 和 CTrace 溯源操作的收敛首先需要节点中存储了足够的信息供查询，然后基站发出溯源请求，通过节点迭代交互，基站逐步构造出路径并最终定位到数据源节点，整个溯源过程需要消耗较多的时间。

(5) 健壮性。

CTrace 和 FBT 设计时没有考虑应对多节点的协同攻击的问题。CAPTRA 在溯源时采用的结合监听节点进行联合判定的机制提供了一定的安全保证，攻击者需要挟持较多的节点才能干扰这种判定机制，这样遭受协同攻击时溯源操作往往会失败但不至于得到错误的结果。PNM 中密钥机制的引入可以防范多种协同攻击手段，具有较强的安全性。

4 结束语

在无线传感器网络中，为了应对节点发生异常尤其是当节点遭俘获发动 DoS/DDoS 攻击时，需要通过溯源方法重构出攻击路径，找出恶意节点的位置，从而采取进一步的措施防止网络遭到进一步的破坏。考虑无线传感器网络资源受限的因素，设计出一种有效而又低开销的溯源方法是一项困难的工作。文中介绍了几种适用于无线传感器网络的溯源方法，这些方法都有各自的优点同时也存在各自的不足之处。只有根据不同的传感器网络环境和应用，在性能和开销之间找到一个平衡点，选择合适的溯源方法，才能得到较好的实践结果。

参考文献：

[1] Culler D, Estrin D, Srivastava M. Overview of Sensor Networks [J]. IEEE Computer Magazine, 2004, 37(8): 41 - 49.

[2] Wood A, Stankovic J. Denial of Service in Sensor Networks [J]. IEEE Computer Society Press, 2002, 35(10): 54 - 62.

[3] AusCERT. AA - 2004.02 - Denial of Service Vulnerability in IEEE 802.11 Wireless Devices [EB/OL]. 2004. http://www.auscert.org.

[4] Xu Wenyuan, Ma Ke, Wade T. Jamming Sensor Networks: Attack and Defense Strategies [J]. IEEE Network, 2006, 20(3): 41 - 47.

[5] Alex C. Hash - Based IP Traceback [J]. ACM SIGCOMM, 2001, 31(4): 3 - 14.

[6] Bellovin S. ICMP traceback messages [S]. Internet Draft: draft - bellovin - itrace - 00.txt, 2000.

[7] Savage S, Wetherall D, Karlin A. Practical network support for IP traceback [J]. ACM SIGCOMM, 2000, 30(4): 295 - 306.

[8] Song D, Perrig A. Advanced and authenticated marking schemes for IP traceback [J]. IEEE INFOCOM, 2001, 2: 878 - 886.

[9] Dong Q, Adler M, Banerjee S. Efficient probabilistic packet marking [C] //IEEE ICNP. [s.l.]: [s.n.], 2005.

[10] Denh S, Lichun Bao. CAPTRA: Coodinated Packet Traceback [C] //In Proceedings of the fifth international conference on Information Processing in sensor networks. [s.l.]: [s.n.], 2006.

[11] Broder A, Mitzenmacher M. Network applications of Bloom filters: a survey [C] //In Proceedings of the 40th Annual Allerton Conference on Communication, Control, and Computing. [s.l.]: [s.n.], 2002.

[12] Zhang Qiyuan, Zhou Xuehai, Yang Feng. Contact - based Traceback in Wireless Sensor Networks [C] //In Proceedings of the International Conference on Wireless Communicaions, Networking and Mobile Computing 2007. [s.l.]: [s.n.], 2007.

[13] Cheng B, Chen H, Liao G. FBT: an efficient traceback scheme in hierarchical wiless sensor network [J]. Security and Communication Networks, 2009, 2: 133 - 144.

[14] Ye F, Yang H, Liu Z. Catching moles in sensor networks [C] //IEEE International Conference on Distributed Computing Systems (ICDCS). [s.l.]: [s.n.], 2007.

(上接第 57 页)

机学报, 2009, 32(5): 862 - 876.

[9] Hand D J, Vinciotti V. Choosing k for two - class nearest neighbor classifiers with unbalance classes [J]. Pattern Recognition Letter, 2003, 24(9): 1555 - 1562.

[10] Cuha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases [C] //In: Hass L M, Tiwary A. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1998: 73 - 84.

[11] Harmer P K, Williams P D, Gunsch G H. An Artificial Immune System Architecture for Computer Security Applications [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3): 252 - 280.

[12] Yang M S, Hu Y J, Lin K C R, et al. Segmenttation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithm [J]. Magnetic Resonance Imaging, 2002(20): 173 - 179.