

一种改进的层次聚类算法

史变霞, 张明新

(西北师范大学 数学与信息科学学院, 甘肃 兰州 730070)

摘 要: 为了更好地实现聚类, 在分析层次聚类(agglomerative)算法和神经网络的 ART2 算法的基础上, 提出了一种改进的层次聚类算法. 改进算法将首先采用一种基于 ART2 的改进神经网络聚类算法得到一个初始的聚类结果, 然后在此基础上利用 agglomerative 算法实现分层聚类. 实验结果表明, 改进算法较原先传统的聚类算法, 不但算法执行速度快、效率高, 而且聚类效果也比较好.

关键词: 层次聚类; 神经网络; ART2

中图分类号: TP391.41

文献标识码: A

文章编号: 1000—7180(2010)12—0055—02

An Improved Algorithm of Hierarchical Clustering

SHI Bian-xia, ZHANG Ming-xin

(School of Mathematics & Information, Northwest Normal University, Lanzhou 730070, China)

Abstract: In order to achieve clustering well, a modified hierarchical clustering algorithm is proposed based on the strengths and weaknesses of hierarchical clustering(agglomerative) algorithm and neural network ART2 algorithm. Improved algorithm will first use an improved ART2 clustering algorithm to form initial clustering results, and then achieve hierarchical clustering result by agglomerative clustering algorithm based on the results of the previous. It is proved that the proposed algorithm is not only faster than the traditional clustering algorithm, but also the clustering result is better.

Key words: hierarchical clustering; neural network; ART2

1 引言

聚类技术在图像处理、知识发现、文档分类和网络入侵检测等领域有着极为广泛的应用^[1-2]. 聚类就是将数据对象分组成多个类或者簇, 划分的原则是在同一个簇中的对象之间具有较高的相似度, 而不同簇中的对象差别较大. 属于一种无指导的学习方法. 目前在文献中存在大量的聚类算法^[3-4]. 算法的选择取决于数据的类型、聚类的目的和应用.

自适应谐振神经网络(Adaptive Resonance Theory Network, ART)^[5-6]是优秀的聚类算法之一, 可以有效实现记忆和学习的平衡、动态聚类, 又具备快速学习能力. 但由于 ART2 采用“winner take all”的竞争规则, 一般只考虑获胜神经元的信

息, 而忽略输出层其他神经元的有用信息. 输出层本质上属于一维或者平面结构, 不能产生层次聚类结果. Agglomerative 层次聚类算法优点在于它可以在不同粒度上对数据进行探测; 然而一旦出现一组对象合并时, 类之间就不能交换对象, 因此在合并前, 必须花费大量时间计算单链接距离(或全链接距离), 并且开始时所合并的簇间距离规模均非常小, 但合并次数却很多, 以致发生在间距很小的簇集之间的最初的聚合迭代却消耗了总时间的 80% 以上^[7]. 针对以上的缺陷, 文中将提出一种综合 ART2 算法和 Agglomerative 聚类算法的分层聚类算法, 在保证聚类效果的前提下有效减少聚类时间.

收稿日期: 2009—11—12; 修回日期: 2010—01—12

基金项目: 国家自然科学基金项目(60673087); 甘肃省自然科学基金项目(3ZS051—A25—047)

2 算法实现

整个算法大致可以分为两个主要步骤:(1)采用一种基于 ART2 的改进神经网络聚类算法得到一个初始的聚类结果;(2)利用 Agglomerative 聚类在初始聚类结果的基础上进行分层聚类.

2.1 初始聚类算法

ART2 算法是神经网络中一种可以对模拟输入信号或二值信号进行无监督聚类的算法,所以 ART2 算法能够降低数据挖掘中原始数据预处理的复杂度,提高挖掘效率.为了克服聚类中心偏移的缺点,本文采用文献[8]的改进算法,是通过比较输入模式与聚类中心的相似度的基础上划分输入模式的类别的,使聚类结果的正确率大大提高.具体算法如下所示.

算法 1:

首先假设 ART2 网络每个新的类的第一个输入为 每一个新类的聚类中心 $\{s_1(1), s_2(1), \dots, s_k(1)\}$. 其中 k 的取值是由每次聚类结果来确定的.

算法的基本步骤:

(1) $z_i = x_i + au_i$

(2) $q_i = \frac{z_i}{e + \|z\|}$

(3) $v_i = f(q_i) + bf(s_i)$

(4) $u_i = \frac{v_i}{e + \|v\|}$

(5) $p_i = u_i + \sum_{i=0}^{m-1} g(y_i)w'_{ij}$

(6) $s_i = \frac{p_i}{e + \|p\|}$

(7) 重复(1)~(6)直到 F_1 层稳定.

(8) $r_i = \frac{u_i + cp_i}{e + \|u\| + c\|p\|}$

当 $\|r\| \geq \rho \cdot e$ 时(ρ 为预先设定的警戒值),则执行(9),否则令 $y_l = 0$,返回(1).

(9) 计算 x_i 到聚类中心的距离,如果: $\|x_i - s_j(1)\| < \|x_i - s_i(1)\|$, 则 $x_i \in J$.

(10) 计算新的聚类中心:

$s_j(l+1) = \frac{1}{N_{j, x_i \in J}} \sum x_i$

(11) 再把 $s_j(l+1)$ 作为 ART2 网络的新的输入,执行步骤(1)~(8).

其中

$f(x) = \begin{cases} x, x \geq \theta \\ 0, x \leq \theta \end{cases}, g(y_i) = \begin{cases} d, j = J \\ 0, j \neq J \end{cases}$

$y_j = \max_j \left\{ \sum_{i=0}^{N-1} p_i w_{ji} \right\}$

$w_{ji} = \frac{u_i}{1-d}, w'_{ij} = \frac{u_i}{1-d}$

2.2 Agglomerative 聚类算法

由算法 1 得到初始聚类结果 $C = \{C_1, C_2, \dots, C_k\}$ 以及聚类中心 $s = \{s_1, s_2, \dots, s_k\}$, 接下来将递归地进行层次聚类. 其基本原理是根据距离合并簇, 其中簇间的距离是依靠计算簇中心之间的距离确定的. $d(x_i, x_j)$ 定义簇 i 和簇 j 之间的对称距离, 即 $d_{ij} = d_{ji}$, 故只需存储上三角阵即可. x_i 表示 i 的特征向量, 簇用中心向量表示. Agglomerative 算法的每一次迭代将距离最小的两个簇 i, j 合并为一个新的簇 m , 并重新计算新簇 m 与其他簇之间的距离. 算法主要步骤如下所示.

算法 2:

(1) for each $(i, j) \in C_2$ 计算簇之间的距离 $d_{ij} = d(s_i, s_j)$;

(2) $(i^*, j^*) = \arg \min(d_{ij}), (i, j) \in C_2$

(3) Set $C_m \leftarrow C_i^* \cup C_j^*$

(4) $C \leftarrow \{C - \{i^*\} - \{j^*\}\} \cup \{m\}$

(5) 中心调整 $s_m = \frac{1}{\|C_m\|} \sum_{v \in C_m} x_v$

(6) 重复(1)~(5)直到所有对象聚类到一个单一的聚类中为止.

3 实验结果

接下来对规模较大的数据进行实验以进一步说明改进的层次聚类算法的有效性. 从 CPU 的运行时间和错误分类率两个方面对它们的运行结果进行比较, 说明改进的算法不仅是有效可行的, 而且效率更高. 表 1 是实验结果的数值对照. 在进行 ART2 神经网络的训练和仿真的时候参数选择 $\rho = 0.92$.

表 1 实验结果对照

数据维数	数据个数	传统层次聚类算法的 CPU 耗时/s	传统算法的错误分类率/%	改进算法的 CPU 耗时/s	改进算法的错误分类率/%
10	100	0.281	13	0.260	12
20	200	3.543	11.5	1.468	11.5
40	500	55.339	12.2	32.654	12.1

4 结束语

层次聚类算法是一种广泛应用的聚类算法, 分
(下转第 61 页)

表 2 处理速度对比

图像大小	结构	平台	时钟数	处理性能/ (幅/s)	处理时间/ ms	工作频率/ MHz
512×512×8 bit	串行结构 ^[5]	TMS320C6204 DSP	2 896 192	69	—	—
	本文结构	Spartan3 xc3s1000 FPGA	59 769	836	1.20	50
1024×1024×8 bit	并行结构 ^[6]	Altera SP1S25 FPGA	—	—	18.18	45
	本文结构	Spartan3 xc3s1000 FPGA	350 449	142	7.01	50

5 结束语

本文提出了一种新的二维离散小波变换结构,完成一层二维离散小波变换只需对存储器读写一次.算法兼顾速度和吞吐率的考虑,采用内部并行计算,相比传统二维小波变换,吞吐率提高了2倍,同时硬件利用率也达到了100%.算法结构灵活,可以配置成多级流水线.

提出的二维小波变换结构,可应用于嵌入式设备、移动设备等实时图像编码系统中,为实时压缩编码提供了有利条件.

参考文献:

[1] ISO/IEC JTC 1/SC 29/WG 1. JPEG 2000 part I final committee draft version 1.0[S/OL]. [2009-09-01]. <http://www.jpeg.org/public/fcd15444-1.pdf>.
[2] 王向阳,崔永瑞,杨红颖,等.基于自适应小波变换的嵌入图像压缩算法[J].微电子学与计算机,2005(2):122-124.
[3] Stephane G, Mallat A. Theory for multiresolution signal

decomposition: the wavelet representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7):674-693.
[4] Tinku Acharya, Chaitali Chakrabarti. A survey on lifting-based discrete wavelet transform architectures[J]. Journal of VLSI Signal Processing, 2006, 42:321-339.
[5] 兰旭光,郑南宁,杨魁志,等. JPEG2000 并行阵列式小波滤波器的 VLSI 结构设计[J]. 电子学报, 2004, 32(11): 47-50.
[6] 孙少林,柴春领,张兆东,等.两种基于独立块的二维提升小波算法在 TMS320C6204 平台上的实现[J]. 军事通信技术, 2006, 27(4):45-48.

作者简介:

曹 辉 男,(1984-),硕士研究生.研究方向为数字图像处理, SOC 设计.
卢 强 男,工程师.研究方向为嵌入式计算机系统结构、数字图像处理、SOC 设计.
黄士坦 男,研究员,博士生导师.研究方向为嵌入式计算机系统结构、高性能计算、数字图像处理、SOC 设计.

(上接第 56 页)

层结构的索引有助于检索及浏览.但是传统凝聚算法的计算性复杂,资源消耗多.为了减少存储空间和计算时间量,本文提出了一种先利用神经网络的方法对数据进行预处理得出初始聚类结果,从而进行聚类,实验表明改进算法有效可行.

参考文献:

[1] Sambasivam S, Theodosopoulos N. Advanced data clustering methods of mining web documents[J]. Issues in Informing Science and Informaion Technology, 2006, 8 (3):563-579.
[2] 马晓春,高翔,高德远.聚类分析在入侵检测系统中的应用研究[J].微电子学与计算机,2005,22(4):134-136.
[3] 李朝健,李朝鹏,李肯立.基于最小生成树的并行分层聚类算法[J].微电子学与计算机,2008,25(9):196-198.
[4] 钱晓凌,王金梁.基于改进 ART2 算法的数据聚类方法

[J]. 兰州交通大学学报,2008,27(6):119-123.
[5] Carpenter G A, Grossberg S. ART2: stable self-organization of pattern recognition codes for analog input patterns[C]// Proceedings of the 1st International Conference on Neural Networks, San Diego. New York: IEEE, 1987.
[6] Kantardzic M. Data mining concepts, models, methods and algorithms[M]. 北京:清华大学出版社,2003.
[7] 李朝鹏,李肯立.基于分层聚类的并行数据预处理算法[J].微电子学与计算机,2007, 24(10):187-189.
[8] 段霞霞,刘彦明.基于 ART2 改进算法的故障聚类研究[J].计算机工程与应用,2008, 44(15):56-57.

作者简介:

史变霞 女,(1980-),硕士研究生.研究方向为数据挖掘、图像处理.
张明新 男,(1962-),博士,教授.研究方向为数据库与数据挖掘、视频图像检索、嵌入式.