Matlab在蚁群聚类算法数据源产生中的应用

陈寿文[1,2],李明东[1]
(1.西华师范大学 微机应用研究所,四川 南充 637000;
2.滁州学院 数学研究所,安徽 滁州 239000)

摘 要:从蚁群聚类算法模拟实验出发,结合待聚类数据源应满足可分性、稳定性和可变性特点的需求,针对数据源的产生和存储问题,给出了一种以服从正态分布的随机向量组来模拟的方法。通过Matlab随机数的产生和数据库的存储与更新技术的阐述,使用不同分布特征下的向量组来抽象数据源的可分性、对向量组的存储及更改,使其满足稳定性和可变性的处理方式,达到了模拟蚁群聚类实验的目的,于.Net环境下用C# Windows应用程序加以实现。在LF算法对产生的随机向量组作用后,运行结果表明该方法产生的数据源是有效的。

关键词:随机数产生;蚁群聚类;.net;C#
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)07-0216-04

Application in Data Source's Generation of Ant Colony Clustering Algorithm with Matlab

CHEN Shou-wen[1,2], LI Ming-dong[1]
(1. Institute of Computer Application, West China Normal University, Nanchong 637000, China;
2. Institute of Mathematical Research, Chuzhou University, Chuzhou 239000, China)

Abstract: Considering the simulation test about ant colony approach for clustering, people often adopt an idea using several groups random vectors obeyed by normal distribution to replace those objects operated. As the data source is satisfied with separability, stability and changeability, these vectors must be met those accordingly. This topic describes a method based on Matlab random numbers' generation to get several groups of stochastic vectors, also, it represents how to save them with SQL server 2005 and recounts how to realize these with csharp windows application based on dot net frameworks. After LF algorithm's action on the vectors, the result of the experiment indicates that the method provides the data source for ant colony algorithm's simulation effectively.

Key words: random numbers' generation; ant colony clustering; dot net; csharp

0 引 言

Matlab是由美国MathWorks公司推出的用于数值计算和图形处理的计算系统环境,它除了具备卓越的数值计算能力外,还提供了专业水平的符号计算、文字处理、可视化建模仿真和实时控制等功能[1]。2008年3月份发布的Matlab R2008a扩展了其Compiler对Visual C++ 2005的支持。基于.Net[2]是各种应用的主流平台,网络程序员可以通过使用Matlab Builder for .Net创建Matlab函数的动态链接库(生成dll文件),并通过.Net环境下C# Windows应用程序的友好界面来实现对库中函数的调用,从而为C#实现蚁群算法提供技术支持。

1 蚁群聚类数据源模拟

在蚁群聚类算法对高维对象进行聚类的过程中,习惯性采用将高维数据对象随机投影到二维网格中[3~5],通过移动的蚂蚁对各聚类对象在其邻域相似度的度量来判断是否对当前对象的移动,最终实现数据聚类的效果。为了验证蚁群算法聚类的效果,产生一些具有分类特性的数据源是前提,而满足不同参数分布的随机向量组能产生明显分类特征,考虑到高级程序语言随机数的产生缺乏灵活性[6]、Matlab产生满足一定分布特征的随机数相当便利两因素,蚁群聚类算法中待聚类的数据源可用Matlab产生的具有明显

收稿日期:2008-10-21;**修回日期**:2009-01-08
基金项目:四川省重点软科学项目{M科技[2006]10号.2006R16-110}
作者简介:陈寿文(1979-),男,安徽怀宁人,硕士研究生,讲师,研究方向为数据挖掘、MIS研发等;李明东,教授,硕士研究生导师,研究方向为MIS研发、数据挖掘。

分类特性的随机向量组来替代,进一步将其随机投影到二维网格中,在实施蚁群聚类算法后即得到聚类结果,从而达到模拟算法的目的。算法运行中参数设置的不同对聚类结果可能产生影响,为了分析二者之间的关系,多次运行算法是必需的,同样,保持待聚类数据源的同一性是先决条件。为此,随机向量在经过 Matlab 相关函数产生后其存储问题需要解决,文中采用将其存储到 SQL Server 数据库中,以便形成固定的数据源,为实验分析提供保障。当需要更换聚类对象时,通过产生新分布特征的向量并及时更改数据库即可,便于为验证、改进蚁群聚类算法做前期准备。

由于高维向量产生机理与二维产生类似,文中用二维向量为例,结合待聚类数据源的产生与存储问题,阐述了 Matlab 产生二维向量组的形成过程,其中向量的各分量为服从 $N(E,\sigma^2)$ 分布的随机数。接下来对 Matlab 随机数的产生做以阐述。

2 Matlab 随机数产生

2.1 Matlab 随机数的类型

Matlab 随机数发生器的种类丰富且用法简便,它不仅包括能生成服从均匀分布的随机数,还包括能生成常用分布(如泊松分布,指数分布等)的随机数。这些随机数发生器所采用的算法,都是经过反复测试并商品化的,其可靠性和稳定性都很强[6]。

2.2 Matlab 随机数的产生

Matlab 中随机数的产生一般采用两种方式:①通过 rand()实现;②通过 randn()实现。对于前者,该函数产生的元素服从[0,1]均匀分布,其取值可以为 $[2^{-53},1-2^{-53}]$之间所有的浮点型数据,这是一般高级语言随机数发生器难以实现的。而后者将产生服从标准正态分布的随机数序列或者是数组,使用方式和前者一致。其他特殊形式的随机数的产生也是在上二者的基础上实现的[7]。

2.3 二维点的坐标

对于数据源中的各二维向量使用平面上点来描述,并采用有序序偶来表示。定义 $G_j=\{(x_i,y_i)\mid x_i\sim N(E_j,\sigma_j^2),y_i\sim N(E'_j,\sigma'^2_j),i=1,\cdots,n\}$ 代表一个类,各维分量服从指定参数下的正态分布,类中共有 n 个点。

2.4 二维点坐标产生

文中采用上述第二种方式来产生随机数,给定一组参数 $(E_i,E'_i,\sigma_i^2,\sigma'^2_i)$,利用 Matlab 下两语句:

$X_j = E_j + \sigma_j^*$ randn(n,1)

$Y_j = E'_j + \sigma'^*_j$ randn(n,1)

产生的随机数组分别作为 G_j 中各序偶(x_{ji},y_{ji})的第 1 维、第 2 维集合,其中 $i=1,\cdots,n$,则

$X_j=\{x_{j1},x_{j2},\cdots,x_{jn}\}$, $Y_j=\{y_{j1},y_{j2},\cdots,y_{jn}\}$

3 数据源的产生和存储

3.1 数据源的产生

利用 2.4 节二维点坐标产生机理,使用 1 组参数值 $(E_i,E'_i,\sigma_i^2,\sigma'^2_i)$ 以生成 G_i,文中采用 4 组不同参数值来获得最终的数据源 $G=\{G_1,G_2,G_3,G_4\}$,概括 G 的特征:$4n$ 个二维点集合,共分成 4 族,每族均有 n 个。

通过 C# 与 Matlab 混合编程方式来实现生成过程,其具体步骤为:

Step1:建立一 NPoint.m 文件,实现返回一 $S_{n\times Num}$ 型矩阵,其中 Num 为族的数目,矩阵中位于(i,j)处的元素值为 Matlab 产生的服从 $N(E_i,\sigma_i^2)$ 分布的随机数;

Step2:利用 Matlab 里 deploytool 工具中 Builder for .net 创建相应项目,生成对应的 dll 文件;

Step3:创建 C# Windows 应用程序项目并做好参数传递界面,将该 dll 文件添加到应用程序中;

Step4:利用 MATLAB.NET.Arrays.MWArray 类型变量接受新添加类实例的返回结果;

其中,NPoint.m 文件如下:

```
function F = NPoint(a,b,num)
%F 为返回值,num 为参数(E_i,E'_i,σ_i^2,σ'_i^2)下随机数个数,利用 a 数组接受各均值 E_i,b 数组接受各平方差 σ_i^2
for i=1:2:length(a)
x=a(i)+sqrt(b(i))*randn(num,1);
y=a(i+1)+sqrt(b(i+1))*randn(num,1);
B=[x y];
if(i==1)
F=B;
else
F=[F B];
end
end
end
```

Myplot.m 文件为:

```
function Myplot(B)
a=size(B);%获取矩阵 B 的行和列数
for i=1:2:a(2) %a(2)为 B 的列数
plot(B(:,i),B(:,i+1),'ro')
hold on
end
```

end

至此,利用 Matlab 随机数发生器产生的二维点坐标集合已经在 C# Windows 应用程序中,接下来将考虑如何实现此数据集存储到 SQL Server 数据库中。与此同时,为了便于验证该数据源的族特征,程序中采用了 Matlab 绘图指令将各 $G_{j\,j=1\cdots4}$ 绘制于同一坐标系中来比照,此功能的实现也是通过 Build for .net 集成在上面同一 dll 文件中,对应的 m 文件为 Myplot.m。

3.2 数据源的存储

在 C# 应用程序中,对返回结果进行数据类型转换后利用数据库相关命令将数据存储到数据库中,以作为稳定的数据源供以后使用。此过程中要注意 Matlab 函数操作及返回的数据均为矩阵类型,在应用程序中只有经过类型转换之后才可以成功实现类成员函数的调用。为此,笔者给出 C# 数值数组和 Matlab 数组转换方式[8]:

①由 C# 到 Matlab 数组转换。

Step1:定义 C# 数组 S1 并对其进行赋值;

Step2:定义 MWNumericArray 型变量 S2,将 S1 强制转换为 MWNumericArray 类型变量并赋值给 S2;

则 S2 可以被 Matlab 函数调用。

②由 Matlab 到 C# 数组转换。

Step1:使用 MWArray 型变量 S3 接受 Matlab 函数返回值;

Step2:将 S3 数组强制转换为 MWNumericArray 类型变量并赋给该类型变量 S4;

Step3:定义 C# 二维数组 S5,将 S4 强制转换为该二维数组并赋值给 S5;

其中 MWNumericArray 是 MWArray 和 C# 中数据的中间类,Matlab 中的数组类型为 MWArray。

4 实验及结果

本实验的运行环境为 Visual Stadio 2005 C# Windows 应用程序,机器为 P4 1.7GHz 526MB 内存,采用 Window XP 操作系统。以二维向量 (x_i, y_i) 满足下述分布特征的 4 族数据为例,实现相应的数据源获取与存储结果,其中:

$G_1 = \{(x_i, y_i) \mid x_i \in N(0,1), y_i \in N(4,1), i = 1,\cdots,100\}$

$G_2 = \{(x_i, y_i) \mid x_i \in N(8,1), y_i \in N(12,1), i = 1,\cdots,100\}$

$G_3 = \{(x_i, y_i) \mid x_i \in N(4,1), y_i \in N(8,1), i = 1,\cdots,100\}$

$G_4 = \{(x_i, y_i) \mid x_i \in N(8,1), y_i \in N(4,1), i = 1,\cdots,100\}$

文中实现数据源 G 的产生并存储的过程均在 C# Windows 应用程序中进行。在此程序中,需要注意对上面生成接口类(dll 文件对应)的使用,在其实例的构造中对应传输参数类型必须正确转换,转换过程参考 3.2 节内容。

程序运行后的结果可以通过文中的几幅图加以说明,其中图 1 为生成的二维点坐标(仅抓取了部分),图 2 为数据存储于数据库中的结果,图 3 显示了各点在坐标系中的散列图象,其中横轴 x 为二维点的横坐标值,y 为该点的纵坐标值。

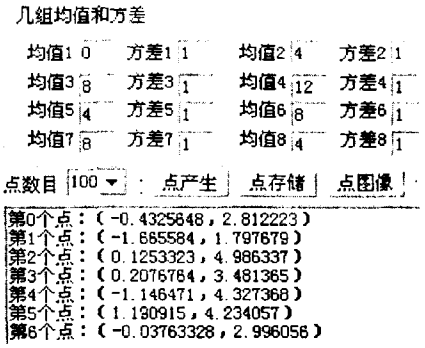

图 1 二维点坐标

表 - dbo.ErWeiXiangliang 摘

ID	X	Y
1	-0.4325648	2.812223
2	-1.665584	1.797679
3	0.1253323	4.986337
4	0.2876764	3.481365
5	-1.146471	4.327368
6	1.190915	4.234057
7	1.189164	4.021466
8	-0.03763328	2.996056

图 2 坐标存于数据库

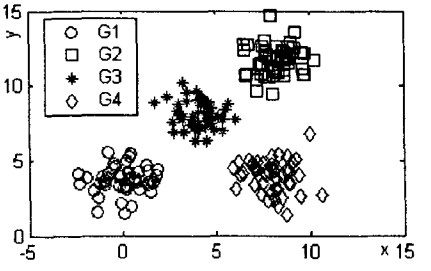

图 3 各点散列图

通过 3.1 节 NPoint.m 代码与图 1 所示,可以观察到该程序是一个动态接受数据的随机向量产生模型。向量维数的增长则只需通过数据库的重新设计来实现,各分量的分布类型的更改可通过修改 NPoint.m 文件中 X 与 Y 的产生函数来实现,程序设计具有通用性特点。

利用蚁群聚类算法 LF[9] 对上数据源进行测试,当采用参数:蚂蚁数目 40,迭代次数 40,$\alpha = 0.2, k_1 = 1.3, k_2 = 1.4$ 时,相应的聚类结果如图 4 所示,其中图 4 横轴 x 为 LF 算法中网格内二维点的横坐标值,纵轴 y 为该点的纵坐标值。由图 4 可以看出,采用 LF 算法进行聚类其效果并不佳,它将数据源中 4 类数据划分成了 3 类,这主要由蚂蚁移动的随机性所致,为此有必要考虑算法的改进问题,此即为下一步目标。

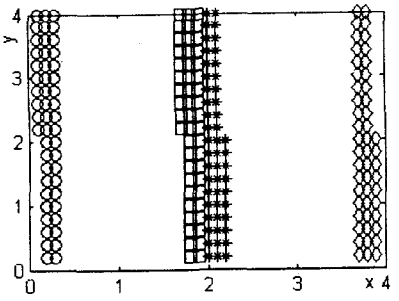

图 4　LF 作用后

5　结束语

从蚁群聚类数据源产生需求出发,利用 Matlab 当前最新版 R2008a 对 .Net 的强大功能支持,结合 C# 被广泛应用的背景,通过二者的混合编程方式实现了随机向量的动态生成和存储,并通过 LF 算法对产生的数据源 G 进行处理,从而说明该方法产生的数据源是有效的。其为蚁群聚类算法改进提供了数据源,也为随机向量的产生提供了一种适用设计方法。

参考文献:

[1] 孟繁娟,杜永平. VB 和 Matlab 混合编程方法 - MatrixVB [J]. 计算机技术与发展,2008,18(5):76 - 78.

[2] 熊　凌. 基于.net 的 Matlab 网络动态数据交换的研究与实现[J]. 微计算机信息,2005,20(8):31 - 33.

[3] 张建华,赵东东,江　贺,等. 一种基于信息素的蚁群聚类算法[J]. 计算机工程与应用,2006,42(3):157 - 159.

[4] 赵伟丽,孙艳蕊,张志国,等. 基于信息熵的蚁群聚类算法的改进[J]. 沈阳化工学院学报,2005,19(4):296 - 300.

[5] Kanade P M, Hall L O. Fuzzy Ants as a Clustering Concept [C]// NAFIPS 2003,22nd International Conference Proceedings. Chicago: North American Fuzzy Information Processing Society,2003:227 - 232.

[6] 马荣国,李平凡. Matlab 接口技术及仿真随机数的生成[J]. 长安大学学报,2003,23(6):111 - 114.

[7] 朱晓玲,姜　浩. 任意概率分布的伪随机数研究和实现[J]. 计算机技术与发展,2007,17(12):116 - 118.

[8] Phan J. Matlab C# Book[M]. America:Colorado State University LePhan Publishing, 2004.

[9] Vizine A, Decastro L, Hruschka E, et al. Towards improving clustering ants: An adaptive ant clustering algorithm[J]. Informatica,2005,29(2):143 - 154.

(上接第 212 页)

[3] Wolfson Microelectronics plc. AC'97 Audio and Touchpanel CODEC WM9712L[S]. Edinburgh, United Kingdom: [s. n.],2003.

[4] 於少峰,严菊明,胡　晨. 基于 AC97 标准的嵌入式音频系统设计与实现[J]. 电子器件, 2004,27(4):733 - 736.

[5] 凌　杰,曹　强. 基于 IntelXscale 的嵌入式音频系统的研究与实现[J]. 嵌入式应用系统, 2006,5(2):11 - 13.

[6] Rubini A. Linux Device Drivers 3rd Edition[S]. Sebastopol, Calif., USA:Oreilly Press, 2005.

[7] 李英伦,杜　旭,项　杰. 基于 IntelXscale 和嵌入式 Linux 的视频模块设计与开发[J]. 计算机工程与设计, 2005,26(9):2282 - 2284.

[8] 肖文鹏. Linux 音频编程指南[S/OL]. 2004 - 02. http://www. ibm. com/developerworks/cn/linux/l - audio/index. html.

[9] 黄晓峰. 多功能芯片驱动设计与实现[D]. 武汉:华中科技大学, 2004.

(上接第 215 页)

[J]. 天津科技大学学报,2004,19(1):11 - 14.

[2] 付保川,班建民,陆卫忠. 基于嵌入式 Web 的远程监控系统设计[J]. 微计算机信息,2005,21(7):58 - 60.

[3] 马忠梅,徐英慧,叶勇建,等. AT91 系列 ARM 核为控制器结构与开发[M]. 北京:北京航空航天大学出版社,2003.

[4] 周立功. ARM 嵌入式系统软件开发实例(一)[M]. 北京:北京航空航天大学出版社,2005.

[5] 李善姬,朴相范,金弘哲. 粮食温度无线检测系统设计[J]. 微计算机信息,2005,21(10):93 - 94.

[6] 王田苗. 嵌入式系统设计与实例开发[M]. 北京:清华大学出版社,2003.

[7] 周小兵. 嵌入式系统 Internet 方案的设计与实现[D]. 成都:电子科技大学,2004.

[8] 罗淳榕,秦现生,马新刚. 基于 CGI 的嵌入式远程控制系统[J]. 测控技术,2006,25(8):50 - 52.

[9] 王彦丽,陈　明,陈　峰,等. 基于 Web Services 企业应用集成的设计与分析[J]. 计算机技术与发展,2008,18(9):212 - 215.

[10] 韩树人,周贤娟. 基于嵌入式 Web 服务器的远程实时数据采集[J]. 计算机技术与发展,2008,18(1):206 - 209.