基于粒子群的近邻传播算法①

谢文斌, 童 楠, 王忠秋, 贾官洪, 陈维奇, 符 强

(宁波大学 科学技术学院, 宁波 315212)

摘 要: 针对近邻传播(AP)算法中偏向参数与收敛系数对 AP 算法的聚类效果的局限性的问题, 提出了一种基于粒子群的近邻传播算法(PSO-AP 算法). 通过将 AP 算法中的偏向参数与收敛系数作为粒子, 然后使用粒子群算法来对其进行智能地调整, 进而提高 AP 算法的聚类效果. 实验结果表明, 该算法能有效地解决偏向参数与收敛系数对 AP 算法的聚类效果局限性, 提高了聚类效果与收敛精度.

关键词: 近邻传播聚类; 粒子群优化算法; 偏向参数; 收敛系数

Affinity Propagation Algorithm Based on Particle Swarm Optimization

XIE Wen-Bin, TONG Nan, WANG Zhong-Qiu, JIA Guan-Hong, CHEN Wei-Qi, FU Qiang

(College of Science and Technology, Ningbo University, Ningbo 315212, China)

Abstract: Aiming at the problem that the preference parameter and damping parameter in affinity propagation algorithm have limitations to the result of clustering, this paper puts forward an affinity propagation algorithm which based on particle swarm optimization (PSO-AP). By taking the two parameters in algorithm as a particle, then adjust it Intelligently by particle swarm optimization (PSO) algorithm, and improve the effect of clustering. The results of experiment show that the algorithm has effectively solved the problem, improved the result of clustering and the accuracy of damping.

Key words: affinity propagation; PSO; preference parameters; damping parameter

近邻传播(Affinity Propagation 简称 AP)聚类算法[1]是由 Frey 等人在 Science 上提出的一种新的快速与有效的消息传递聚类算法. 近邻传播算法的优势如下: 第一, 该算法可以对较大的数据集进行较好较快的聚类; 第二, 该算法把每个数据样本作为待聚类的数据中心, 这使其不受初始中心限制; 第三, 该算法对相似度矩阵的对称性没有要求, 这扩展了该算法的应用范围. 由于近邻传播算法的种种优点, 该算法被广泛的应用于人脸识别、手写体字符识别、基因识别、最优航空路线确定等问题上.

该算法中有两个重要的参数[2]: 偏向参数与收敛系数. 这两个参数对聚类结果都有很大的影响: 偏向参数主要影响最后的聚类数目, 收敛系数主要影响算法的收敛速度与精度, 但这两个参数往往需要通过实验分别进行调试取值, 不但过程复杂, 而且难于获得最佳参数值, 在很大程度上局限了 AP 算法的聚类效果.

本文针对上述问题提出了基于粒子群的近邻传播算法, 通过粒子群算法的全局寻优能力来智能调整偏向参数与收敛系数的值, 使 AP 算法实现全局最优聚类效果. 利用新方法对各种类型的数据进行了聚类实验分析, 实验结果验证了算法的有效性.

1 近邻传播聚类

与 K-means、模糊 C 等聚类算法相比较, AP 聚类算法不需要初始中心, 它将每个数据点作为候选的聚类中心, 通过数据点之间的吸引与归属关系进行聚类.

AP 算法是根据建立的相似度(Similarity)矩阵进行聚类的. 相似度矩阵 S 按(1)式建立, 其的非对角线元

① 基金项目:浙江省教育厅科研项目(Y201326770);宁波大学科研基金项目(XYL12009);浙江省教育厅科研项目(Y201326872);浙江省教育科学规划课题(SCG090)

收稿时间:2013-07-23;收到修改稿时间:2013-09-22

素 $s(i,k)$ 为点 x_i 与点 x_k 之间的关系，对角线元素 $s(k,k)$ 为偏向(Preference)参数 $P(k)$，$P(k)$ 的初始值一般取相同的值，为相似度矩阵中所有非对角线元素最小值或均值，其的初始大小对最后的聚类数有较大的影响，P越大产生的聚类个数越多，反之亦然.

$$s(i,k)=\begin{cases}-\|x_i-x_k\| & i\neq k\\ P & i=k\end{cases} \tag{1}$$

AP 算法的核心为数据点之间相互的信息传递, AP 算法有两种信息[3]，它们分别为吸引度(Responsibility)与归属度(Availability)，建立过程如图 1、图 2 所示.

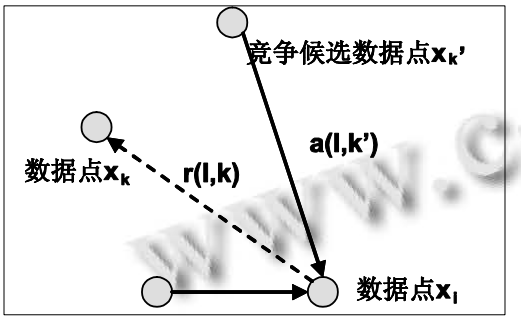


图 1　吸引度的建立

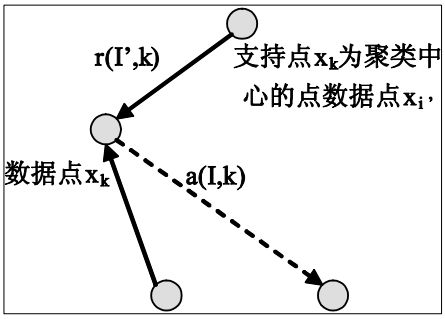


图 2　归属度的建立

算法开始吸引度 $r(i,k)$ 与归属度 $a(i,k)$ 的初值都为 0，表示开始时数据之间没有任何聚类关系，它们按(2)到(6)式更新. $r(i,k)$ 为由点 x_i 传到候选聚类中心点 x_k 的信息，它反映候选聚类中心点 x_k 作为点 x_i 的聚类中心的适应程度. $a(i,k)$ 为由候选聚类中心点 x_k 传到其所有潜在聚类成员点 x_i，它反映点 x_i 作为点 x_k 的聚类成员的适应程度. 其中 $r(k,k)$ 与 $a(k,k)$ 为点 x_k 的自吸引度与自归属度这两个值越大说明其越适合作为聚类中心[4]，一般把 $r(k,k)$ 与 $a(k,k)$ 之和大于 0 的点就认为其是一个较好的聚类中心.

$$r(i,k)=s(i,k)-\max_{k's.t.k'\neq k}\{a(i,k')+s(i,k')\} \tag{2}$$

$$r_{new}(i,k)=\lambda\times r_{old}(i,k)+(1-\lambda)\times r(i,k) \tag{3}$$

$$a(i,k)=\min\{0,r(k,k)+\sum_{i's.t.i'\notin\{i,k\}}\max\{0,r(i',k)\}\} \tag{4}$$

$$a(k,k)=\sum_{i's.t.i'\notin\{i,k\}}\max\{0,r(i',k)\} \tag{5}$$

$$a_{new}(i,k)=\lambda\times a_{old}(i,k)+(1-\lambda)\times a(i,k) \tag{6}$$

其中下标为 old 的代表上一次的结果, new 代表本次更新后结果. λ 为收敛系数($\lambda\in[0,1)$)，λ 越大消除振荡的效果越好，但收敛速度也越慢，反之亦然.

迭代的终止条件为: 迭代次数超过最大值或者当聚类中心连续多少次不发生改变时终止迭代.

最后根据得到的中心结合相似度矩阵对数据点进行分类.

2　实验数据集与指标

本实验中采用了 BWP 指标[5]对聚类结果进行评价. BWP 指标是基于样本的聚类距离和聚类离差距离提出的.

BWP 指标适用于衡量聚类数大于 1 的数据集，该指标值的范围在-1 到 1 之间，该值越大说明聚类结果越好. 之所以选用 BWP 指标，是因为该指标可以有效地反应类内紧密性与类间分离性，并且能对聚类结果进行一个有效的评估.

本文选用了表 1 所示的 5 个数据集作为测试数据集对本文提出算法进行验证和比较. 其中选用的每个数据集均为聚类分析中常用的数据集，而且各具特点，有利于对聚类算法进行一个全面的分析.

表 1　数据集特点

数据集	数据个数	数据维数	数据类数	数据类型
Aggregation1	788	2	7	混合
Aggregation2	195	2	2	紧密、环形
Aggregation3	307	2	2	不完全分离、紧密
Aggregation4	232	2	2	不完全分离、松散
Aggregation5	143	2	6	完全分离、松散

3　AP算法参数分析

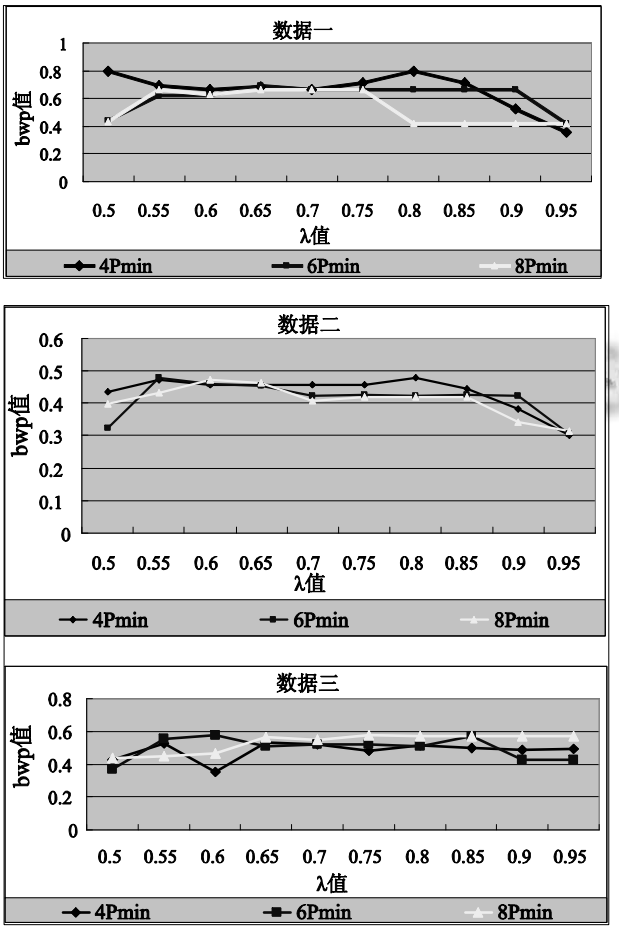
为证明参数调整的必要性，以及通过参数调整优化算法的可行性. 本实验通过传统的方法对 AP 算法的参数进行了分析.

3.1 收敛系数 λ 分析

在传统 AP 算法中收敛系数一般为固定值，但收敛

系数对于 AP 算法的最终结果以及运算过程都存在很大的影响[6], 合适的收敛系数可以保证较好的收敛速度和迭代的稳定性. 对于不同的数据集, 采取不同的收敛系数也会产生影响. 本实验通过对多个不同的数据集进行实验, 以取出一个普遍能应用到处理各种数据集的 AP 算法的收敛系数 λ. 为了验证单调收敛系数 λ 实验的合理性, 需要保证偏向参数 P 不变, 同时为了能够验证实验的普遍性, 我们对于不同的数据集实验采取三个普遍的 P, 分别为 4 倍 Pmin, 6Pmin 以及 8Pmin.

图 3 是对收敛系数 λ 单调实验的结果. 对于不完全分离以及排列紧密的数据集三, 可以观察出取不同的 λ 值对于 bwp 值动荡比较大, 说明处理较难聚类的数据集, λ 值对于实验结果影响很大. 仔细观察在 λ 取 0.8-0.9 时, bwp 值较为稳定, 并且 bwp 值相对较大, 说明聚类效果好. 相反对于完全分离的数据集六, 不同的 λ 值对于实验结果影响不大, 收敛系数对于聚类效果作用很小. 综合以上六个数据集的 λ 单调实验, 我们得出结论, 当 λ 取值为 0.8-0.9 时, bwp 值较为稳定, 并且数值相对较大, 收敛系数取此值对于数据集的聚类效果明显.



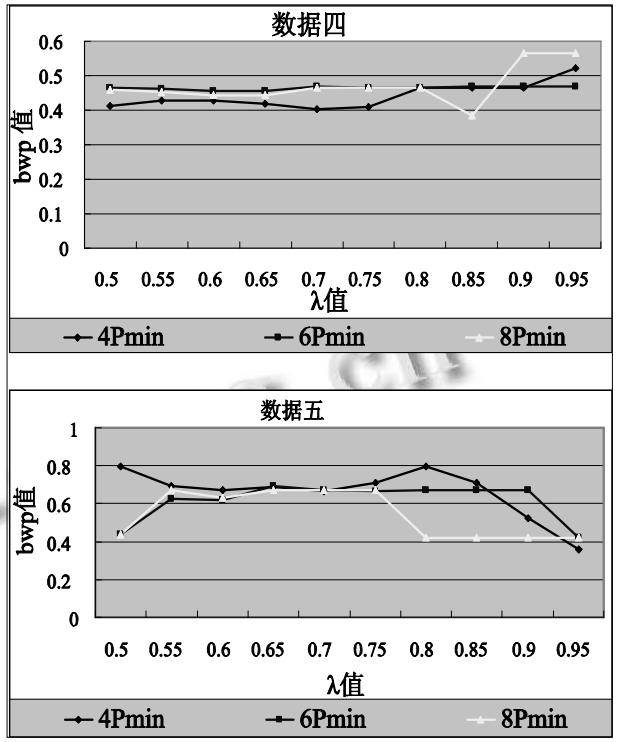
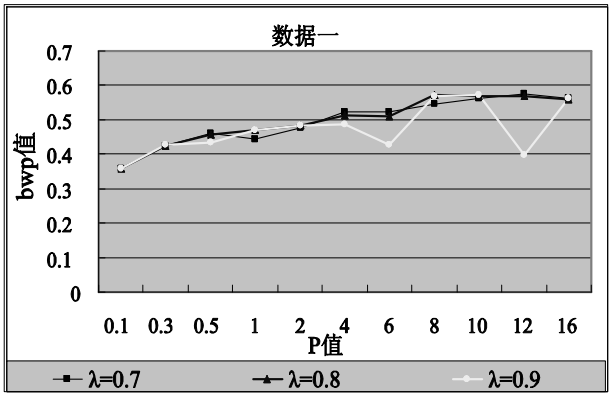


图 3 BWP 指标对单调 λ 的结果对比

3.2 偏向参数 P 分析

偏向参数 P 值[7]为相似度矩阵上的主对角线元素, 其初始取值对于数据集最后的聚类数目有很大的影响. 下面我们单独对 P 的初始值进行调节, 针对 6 个不同类型的数据集, 采用收敛系数 λ 为 0.7、0.8、0.9, P 值分别从 0.1 倍 Pmin 到 16 倍 Pmin 之间取 11 个数值进行试验.

实验结果如图 4 所示, 从图中我们可以对于前四个数据集来说偏向参数的取值在 4~8 倍时比较合适. 但对于数据五却不然, 其在 0.3 倍的时候效果最佳. 4~8 倍只适合于某些数据集, 没有普遍性.



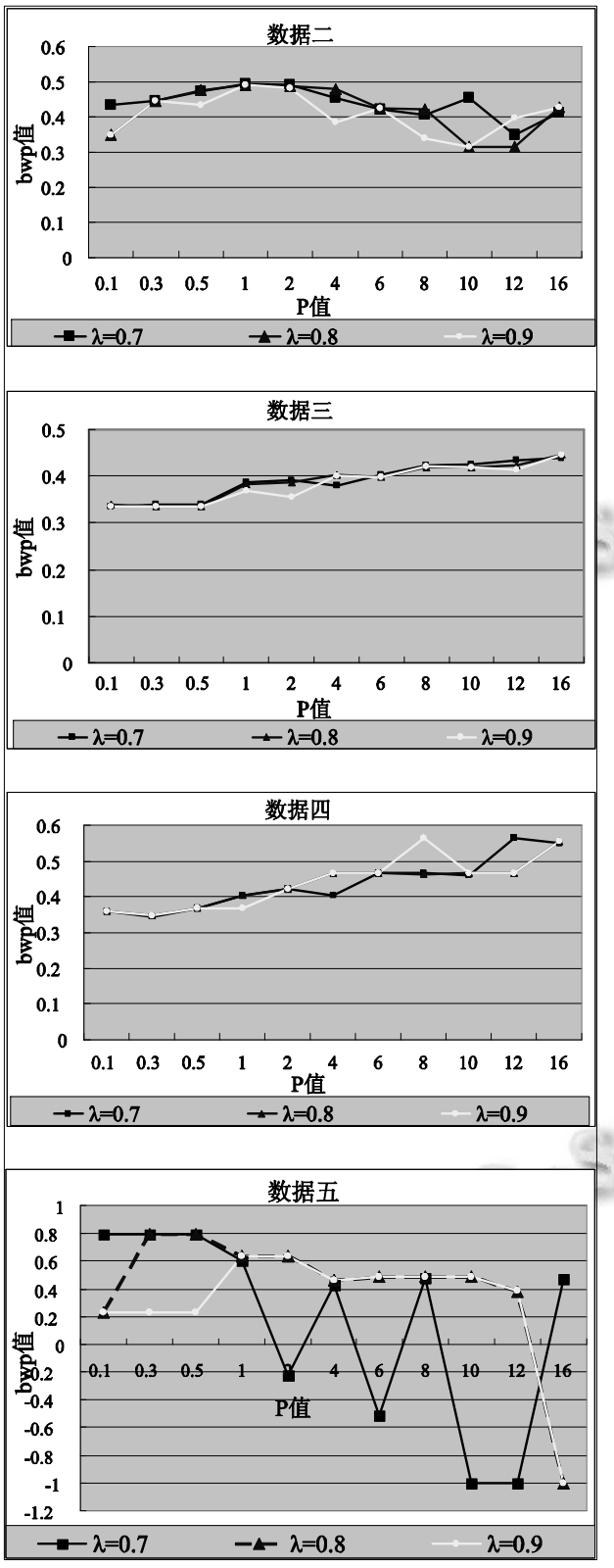


图 4　BWP 指标对单调 P 的结果对比

通过上述实验分别对偏向参数与收敛系数进行调试取值，就这种方法而言其的过程十分复杂，而且难于获得最佳参数值，虽然最后还是得出了一个大概的取值范围，但是在图 4 中的数据五可以发现这个大概的范围并不具有普遍性. 那就使得我们不能使用在一定范围内单独调整参数的方法来获得最优解，我们需要对这个两个参数值进行同时调整与全局搜索. 由此本文题出了出了基于粒子群的近邻传播算法，通过粒子群算法来智能的调整这两个参数的值.

4　基于粒子群的近邻传播算法

4.1 PSO-AP 算法原理及步骤

PSO-AP 算法就是将偏向参数与收敛系数作为粒子群算法[8]中粒子的位置坐标. 初始化各粒子坐标与速度，即选取不同的偏向参数与收敛系数与其初始的变化方向. 然后按照式(7)、(8)不断更新粒子的位置与方向，在更新的过程中将粒子的位置作为 AP 算法的偏向参数与收敛系数的值进行聚类，并用BWP指标衡量聚类效果作为该粒子的适应度.

$$V_{id} = \omega V_{id} + \eta_1 \text{rand}()(P_{id} - X_{id}) + \eta_2 \text{rand}()(P_{gd} - X_{id}) \quad (7)$$

$$X_{id} = X_{id} + V_{id} \quad (8)$$

其中 V_{id} 表示第 i 个粒子在第 d 维上的速度，p_{id} 为该粒子经历过的最好位置，p_{gd} 为群体所经历的最好位置，ω 为惯性权重，η_1、η_2 为调节 p_{id} 和 p_{gd} 相对重要性的参数. (7)式将当前粒子位置与个体最优解与群体最优解对比，得到一个群体最优与个体最优的发展趋势，再根据这个发展趋势与原来初速度的方向确定新的速度方向. (8)式就是在之前的到的向上运动一定距离产生粒子的新位置. 其中 X_{nd} 的奇数维上值为新的偏向参数偶数维上的值为收敛系数，如(9)到(10)式所示.

$$P = X_{nd} \qquad \text{当d为奇数} \quad (9)$$

$$\lambda = X_{nd} \qquad \text{当d为偶数} \quad (10)$$

迭代结束的条件为：当迭代次数超过最大值或者当聚类中心连续多少次不发生改变时终止迭代.

新型 AP 算法(PSO-AP)的实现步骤如下:

Step1: 初始化粒子群，设置各粒子的初始位置(即参数偏向参数与收敛系数的值)与初始速度(即参数偏向参数与收敛系数的值变化的方向);

Step2: 根据(1)式建立相似度矩阵 S，初始化信息矩阵;

Step3: 根据(9)、(10)式更新 AP 算法的参数;

Step4: 根据(2)到(6)式更新信息矩阵;

Step5: 如果达到结束条件，则结束，否则跳回 Step4;

Step6: 计算每个粒子的适应度(即每组参数偏向参数与收敛系数的聚类效果);

Step7: 对比每个粒子，比较它的适应度值和它经历过的最好位置的适应度值 p_{id}，如果更好，则更新 p_{id};

Step8: 对比每个粒子，比较它的适应度值和群体所经历的最好位置的适应度值 p_{gd}，如果更好，则更新 p_{gd};

Step9: 根据(7)式更新每个粒子的速度;

Step10: 根据(8)式得到粒子移动的下一位置;

Step11: 如果达到结束条件，则结束，否则转 Step3.

4.2 实验分析

本实验中 PSO-AP 算法的 η_1、η_2 根据经验均取值为 2. 由于 ω 较大时算法具有较强的全局搜索能力，ω 较小时算法倾向于局部搜索[9]. 所以令 ω 等于(11)式，随着叠代进行，ω 由最大加权因子 $\omega_{\max}$ 减少到最小加权因子 $\omega_{\min}$. 本实验 $\omega_{\max}$ 与 $\omega_{\min}$ 根据经验分别取值为 0.9 与 0.4.

$$\omega = \omega_{\max} - iter\frac{\omega_{\max} - \omega_{\min}}{iter_{\max}} \tag{11}$$

式中 iter 为当前迭代数，$\text{iter}_{\max}$ 为总迭代数.

为更好说明 PSO-AP 算法的优化效果，同时利用粒子群算法与普通的 AP 算法(偏向参数 P 取相似度矩阵的中位值，收敛系数 λ 取 0.85)分别对如表 1 所示的 5 个数据集进行聚类精度测试，实验结果如图 5、表 3 所示. 同时以数据集 4 为例给出了算法的迭代过程，如图 6 所示.

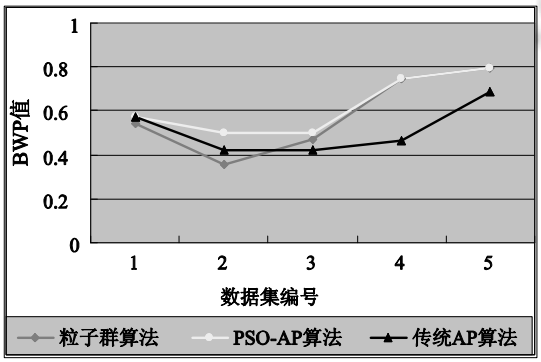


图 5 BWP 指标对聚类算法的性能的评价结果对比

通过图 5 可以明显看出 PSO-AP 算法在聚类精度上要远优于普通的 AP 算法. 图 6 则显示了 PSO-AP 算法实现参数调整，对聚类效果的优化过程.

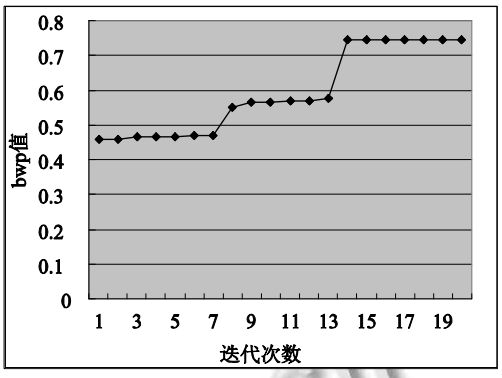


图 6 数据集 4 迭代次数与优化值的曲线图

表 2 各算法运算时间

数据集	PSO-AP 算法(秒)	粒子群算法(秒)	传统 AP 算法(秒)
Aggregation1	58.045	66.317	21.514
Aggregation2	4.455	10.440	0.577
Aggregation3	9.676	12.515	2.264
Aggregation4	3.229	6.190	0.579
Aggregation5	2.228	7.035	0.237

表 3 BWP 指标对聚类算法的性能的评价结果

数据集	数据类型	粒子群算法	PSO-AP 算法	传统 AP 算法
Aggregation1	混合	0.5411	0.5712	0.5693
Aggregation2	紧密、环形	0.3537	0.4972	0.4203
Aggregation3	不完全分离、紧密	0.4689	0.5018	0.4187
Aggregation4	不完全分离、松散	0.7459	0.7459	0.6846
Aggregation5	完全分离、松散	0.7963	0.7963	0.4218

表 2 显示 PSO-AP 算法运算速度要比 AP 算法要慢，这是由于 PSO-AP 算法对 AP 算法的参数进行了调整，以提高算法的聚类精度与适应性，但相比粒子群算法要快得多. 从表 3 可以发现，PSO-AP 算法的聚类效果均明显优于 AP 算法. 而粒子群算法与 PSO-AP 算法对 Aggregation4、Aggregation5 聚类后的结果基本相同，这是因为这些数据集都是比较松散或是完全分离的数据

(下转第 76 页)

(1) 详细的信息反馈机制.

(2) 可靠的传输保证机制(断点续传).

(3) 操作简单, 便捷.

(4) 完善的日志记录, 易于排错.

由于目前大多数企业的文件传输使用的是操作系统自带的FTP功能, 配置复杂, FTP较为分散且不利于管理, 也没有断点续传功能. 由于大多数企业数据还没有达到一定规模, 商业的文件传输系统并没有得到企业的重视, 商用文件传输系统还没有实际应用的环境, 但随着企业不断发展, 数据量不断增大, 大数据时代的逐渐来临, 商业文件传输, 作为客户端与服务端的数据传输工具, 极大地方便了工作, 提升了工作效率. 如果在此基础上对系统进行进一步完善, 构建成复杂大型的文件传输引擎, 将会产生极大的商业价值和经济效益.

参考文献

1 百度百科.CuteFTP 介绍.http://baike.baidu.com/view/177879.htm.2013-3-13.

2 百度百科.LeapFTP 介绍 http://baike.baidu.com/view/119580.htm.2013-3-13.

3 百度百科.FlashFXP 介绍.http://baike.baidu.com/view/177890.htm.2013-3-13.

4 百度百科.Serv-U 介绍.http://baike.baidu.com/view/537933.htm.2013-3-13.

5 王正,罗万明,阎保平.并行下载最优机制.软件学报,2009,20(8).

6 邓湘,吴迪.基于 P2P 的文件并行上载机制研究.http://www.doc88.com/p-995532369499.html. 2013-3-15.

7 志良的技术博客.深入探析 c# Socket.http://www.cnblogs.com/tianzhiliang/archive/2010/09/08/1821623.html.2013-3-15.

8 Palmer G. 康博译.C#程序员参考手册.北京:清华大学出版社,2002.

9 周存杰.Visual C#.NET 网络核心编程.北京:清华大学出版社,2002.

10 曾强聪.软件工程.北京:高等教育出版社,2004.

(上接第 107 页)

集, 聚类起来相对容易, 均能获取较好的聚类效果. 但对比 Aggregation1、Aggregation2 、Aggregation3 这三个较难的数据集来说粒子群算法的聚类精度就明显不如 PSO-AP 算法. 总的来说 PSO-AP 算法在对数据集的适应度与聚类精度上都要优于粒子群算法与普通的 AP 算法.

5 结语

本文针对近邻传播(AP)算法中偏向参数与收敛系数对 AP 算法的聚类效果的局限性的问题进行了分析研究. 首先, 本文通过实验分别对偏向参数与收敛系数进行调试取值, 研究了偏向参数与收敛系数的特性. 然后, 结合偏向参数与收敛系数的特性, 提出了通过粒子群算法搜寻最优的偏向参数与收敛系数的方法, 使近邻传播算法达到一个最佳的聚类结果. 最后, 通过实验证明了 PSO-AP 算法的有效性, 且在聚类精度与适应性上要优于普通的 AP 算法与粒子群算法.

参考文献

1 Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976.

2 王开军,张军英等.自适应仿射传播聚类.自动化学报,2007,33(12):1242–1246.

3 刘胜宇,刘家锋等.基于改进 AP 聚类算法的人脸标注技术研究.智能计算机与应用学报,2011,1(1):35–38.

4 杨传慧,吉根林等.AP 算法在图像聚类中的应用研究.计算机与数字工程学报,2012,40(10):119–121.

5 周世斌,徐振源等.一种基于近邻传播算法的最佳聚类数确定方法.控制与决策学报,2011,26(8):1147–1157.

6 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报,2008,19(11):2803–2813.

7 邢艳,周勇.基于互近邻一致性的近邻传播算法.计算机应用技术,2012,29(7):2524–2526.

8 刘靖明,韩丽川等.基于粒子群的K均值聚类算法.系统工程理论与实践,2005,6:54–58.

9 陶新民,徐晶等.一种改进的粒子群和K均值混合聚类算法.电子与信息学报,2010, 32(1):54–58.