

为了比较样本数量对 PGAK- I 型和 PGAK- II 型染色体并行遗传效率的影响,我们在 20 个二维数据点周围分别产生正态分布的大量样本点,并在这些样本点中随机抽取样本组成多个随机样本点集合,采用 PGAK- I 型和 PGAK- II 型染色体并行遗传算法分别对这些样本集进行聚类,实验所得聚类时间与样本数之间的关系如图 5 所示。

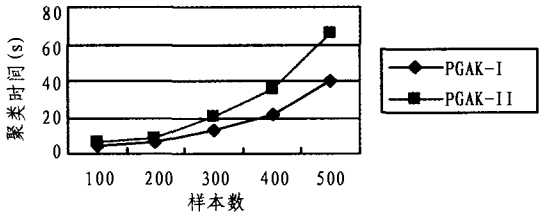


图 5 样本对可变量染色体编码的并行遗传算法的影响

从图 5 可以看出,样本数对 PGAK- I 型染色体的影响远小于 PGAK- II 型染色体,PGAK- I 型染色体在性能上要优于 PGAK- II 型染色体。这是因为 PGAK- II 型染色体中的基因是由聚类中心对应的样本点数据组成,而 PGAK- I 型染色体中的基因则只是由聚类中心对应的样本点编号组成,而且在遗传过程中,染色体长度在动态变化,在样本数增多的时候,染色体长度也会加长。因此 PGAK- II 型染色体占用的内存空间必然比 PGAK- I 型染色体大。随着样本数的增加,这种差别将越来越大,具体的性能差别当然会表现得较为明显。

5.2 算法性能测试

为了验证本文所提出的基于并行遗传算法的 K-means 聚类(PGAKClust)方法的实际性能,我们随机生成 12 个二维样本集,分别使用 K-Means 算法和本文提出的 PGAKClust 聚类方法对上述 12 个样本集进行聚类(采用 PGAK- I 型染色体编码)。聚类结果如表 2 所示。

表 2 算法性能测试结果

样本集编号	样本数	实际类别数	K-Means K=5	K-Means K=10	K-Means K=15	PGAKClust 聚类数	PGAKClust 平均准确率
1	100	4	62%	54%	41%	4	94.8%
2	100	6	61%	57%	43%	6	94.6%
3	100	8	54%	59%	48%	8	94.5%
4	100	10	51%	66%	52%	10	94.6%
5	200	8	53%	57%	46%	8	94.1%
6	200	10	50%	64%	49%	10	94.0%
7	200	12	47%	56%	54%	12	93.8%
8	200	14	41%	51%	59%	14	93.7%
9	500	10	48%	61%	47%	10	93.0%
10	500	12	44%	54%	51%	12	93.1%
11	500	14	38%	49%	55%	14	92.9%
12	500	16	31%	42%	56%	16	92.8%

从表 2 可以发现:对于 K-Means 算法,由于聚类前无法获知样本类别数,只能通过大致估计类别数来进行聚类运算,当估值的聚类数目与实际类别数差别较大时,聚类精确度将显著下降;同时,由于算法对聚类初始中心选取的敏感性,使得聚类结果精确度相对较低;此外,随着样本数的增加和样本类别数的增加,聚类精确度也会明显下降。

对于本文所提出的 PGAKClust 聚类方法,由于采用可变量染色体编码的并行遗传算法,通过进化计算,能动态挖掘样本聚类数目,同时优化了初始聚类中心的选择,因此在避免聚类数盲目估计的同时能获取最优初始聚类中心,从而保证了聚类的精确度;此外,通过并行遗传算法的合理优化,使得样本数和样本类别数对聚类结果的影响得以有效降低。

结束语 本文通过提出一种基于并行遗传算法的 K-means 聚类方法,克服了传统 K-means 算法对初始聚类中心选择的敏感,以及聚类数 K 难以确定的问题,充分发挥了 K-means 算法的高效性和并行遗传算法的全局优化能力。有效地均衡了算法对聚类空间的探索和开发能力,实验证明该算法是一种高精度的聚类方法。

参考文献

[1] Larsen B, Aone C. Fast and effective text mining using linear-time document clustering[A]//Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, August 1999:16-22

[2] MacQueen J B. Some methods for classification and analysis of multivariate observations//Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297

[3] Zhao Y, Karypis G. Criterion Functions for Document Clustering: Experiments and Analysis[R]. Technical Report. # 01-04. Department of Computer Science, University of Minnesota, 2001

[4] Steinbach M, Karypis G, Kumar V. A comparison of Document Clustering Techniques[R]. Department of Comp Sci & Eng University of Minnesota, 2000:1-20

[5] Salton G, Wang A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11): 613-620

[6] Mühlenbein H. Evolution in time and space-the parallel genetic algorithm. In Rawlins, Foundations of Genetic Algorithms, Morgan Kaufmann, 1991

[7] Hung S L, Adeli H. A parallel genetic/neural network learning algorithm for MIMD shared memory machines. IEEE Transactions on Neural Networks, 1994, 5(6):900-908

[8] Liu Juan, Iba H. Selecting informative genes with parallel GA in tissue classification. Genome Informatics, 2001, 3(12): 14-23

[9] Goldberg D E, Deb K, Korb B. Don't Worry Be Messy//Proc. of ICGA. 1991:24-30

[10] Ramze R M, Lelieveldt B P F, Reiber J H C. A new cluster validity indexes for the fuzzy c-mean[J]. Pattern Recognition Letters, 1998, 19:237-246

[11] Makoto I, Takenobu T. Hierarchical Bayesian clustering for automatic text classification[R]. Tech Rep. TR95-0015. Department of Computer Science Tokyo Institute of Technology, 1995

更正

我刊 2008 年 Vol. 35No. 3“基于免疫重构的阴性选择算法”一文中,图 1 更正如右图。由此给读者带来的不便,表示歉意。

《计算机科学》编辑部

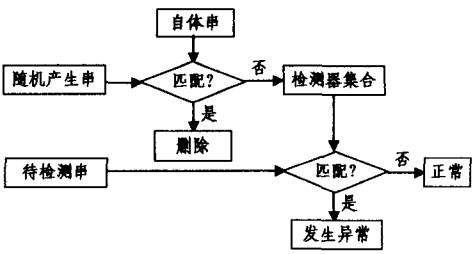


图 1 阴性选择算法流程图