doi:10.3969/j.issn.1673-4785.2009.06.013

支持向量数据描述的基因表达数据聚类方法

季瑞瑞,刘 丁
(西安理工大学 信控中心,陕西 西安 710048)

摘 要:为改善传统的基因表达数据聚类方法正确率偏低的问题,研究了支持向量数据描述(SVDD)算法在基因表达数据聚类中的应用,该方法通过寻找最优分类超球实现对数据集的有效聚类.将类间信息融入聚类有效性评估准则中,通过模拟退火优化算法寻找SVDD算法中的最优核函数参数和惩罚因子,在训练时引入非样本数据提高运算效率.对酵母细胞生长周期的基因表达数据集的仿真实验结果表明,在新的聚类有效性评估准则下进行参数寻优,能够更快更好地得到最佳参数,同时,算法具有聚类精度高和运算速度快的优点.

关键词:基因表达数据;支持向量数据描述;聚类;模拟退火
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-4785(2009)06-0544-05

Improved gene expression data clustering using a support vector domain description algorithm

JI Rui-rui, LIU Ding
(Center of Information and Control Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: The application of the support vector domain description (SVDD) algorithm in gene expression data clustering was proposed as a means to improve the low accuracy of current clustering methods. This method effectivly clustered the dataset by finding the optimal separating hyper-sphere. Inter-class information was introduced into the current clustering assessment criterion in the form of a minimum within-class distance. The simulated annealing (SA) algorithm was used to find the optimal kernel function parameter and the punishment factor of the SVDD algorithm. Non-sample data were added in training to increase computational efficiency. Simulation results using the yeast cell cycle expression dataset showed that optimal parameters can be obtained faster and more accurately with the new assessment criteria. Similar improvements were found in clustering accuracy and speed.

Keywords: gene expression data; SVDD; clustering; simulated annealing

随着人类基因组计划(HGP)的顺利实施与基因芯片技术的发展,人们可以观察到成千上万的基因在某个生命现象中的表达情况.由于生物体本身的复杂性,这些数据往往是高维、海量的,如何从这些数据中挖掘出有用的信息,发现基因的功能具有重要的研究意义.目前对基因表达数据的处理主要是进行聚类分析.常用的聚类算法有K-均值法(K-means)[1]、自组织映射法(SOM)[2-3]、神经网络[4]、主元分析[5]、支持向量机(SVM)[6]、动态模型[7]、隐马尔可夫模型[8]等,其最终目的是寻找多类目标样本集的最佳划分,同一类一般是具有已知功能的基因,这样利用聚类结果可以对未知功能的基因进行划分和识别.

传统的聚类方法虽然能够得到不错的效果,但是存在一定的弊端,如:需要预先指定聚类数目;对边界和噪声数据敏感以及误判问题;如果需要加入新的类别,必然影响整个系统.起源于SVM的支持向量数据描述算法(support vector domain description,SVDD)[9-10]把聚类看作是样本的"认知",通过寻找覆盖样本在特征空间的最优超球实现对数据的聚类,不仅减少了误判率,同时新类别的介入也不需重新训练全部样本.研究了基于SVDD的基因表达数据的聚类问题,改进了聚类有效性评价准则,并以此作为寻找SVDD参数的准则,通过优化算法寻找最佳参数,提高了计算效率,改善了误判问题,从而

收稿日期:2009-06-23.
通信作者:季瑞瑞. E-mail: hezirui423@163.com.

提高了对于未知功能基因的认知能力.

1　支持向量数据描述

支持向量聚类的基本思想是通过在特征空间中寻找包围目标样本点的超球体,并通过最小化该超球体的体积,使得目标样本点尽可能的被包围在超球体内部,而非目标样本点尽可能的在超球体外部,从而实现不同类之间的有效划分. 超球体内的点认为是目标类数据,超球体外的点被认为是非目标类数据,位于球表面上的点就是支持向量. SVDD 采用超球来覆盖样本数据,使得聚类的收敛域更小,聚类效果更精确,从而较好地解决误判的问题.

1.1　问题描述

SVDD 问题描述如下:设有 n 个样本数据,包围这些样本的最小超球的球面中心为 a,球面半径为 R,则寻找该最小超球的过程变成求解以下的目标函数:

$$\min[R^2 + C\sum_i \xi_i], \tag{1}$$

$$\text{s.t.} \ \| \Phi(x_i) - a \|^2 \leqslant R^2 + \xi_i, \xi_i \geqslant 0, \forall i. \tag{2}$$

式中:C 为惩罚因子,ξ_i 是为了提高超球的鲁棒性,允许包含非目标样本而引入的松弛变量.

将上述问题引入 Lagrange 乘子 α_i 和 γ_i,转化成 Lagrange 极值问题:

$$L(R,a,\xi,\alpha,\gamma) = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[R^2 + \xi_i - (x_i \cdot x_i - 2ax_i + a \cdot a)] - \sum_{i=1}^{n}\gamma_i\xi_i.$$

式中:$\alpha_i \geqslant 0, \gamma \geqslant 0$.

求解该式的极值,分别对 R、a 和 ξ_i 求偏导并令其为 0,得

$$\sum_{i=1}^{n}\alpha_i = 1, a = \sum_{i=1}^{n}\alpha_i x_i,$$

$$C - \alpha_i - \gamma_i = 0, \forall i.$$

由于 $\alpha_i \geqslant 0, \gamma_i \geqslant 0$,可以将 $C - \alpha_i - \gamma_i = 0$ 转换成 $0 \leqslant \alpha_i \leqslant C$.

重新改写上面的等式,整个问题变为求解式(3):

$$\max L = \sum_{i=1}^{n}\alpha_i(x_i \cdot x_i) - \sum_{i=1,j=1}^{n}\alpha_i\alpha_j(x_i \cdot x_j), \tag{3}$$

$$\text{s.t.} \sum_{i=1}^{n}\alpha_i = 1, 0 \leqslant \alpha_i \leqslant C. \tag{4}$$

由于实际的样本分布不一定是球形的,根据 Vapnik 的理论[11],利用满足 Mercer 条件的 $K(x_i,x_i)$ 代替内积运算,将样本映射到一个高维特征空间,当选择合适的核函数时,可以得到关于样本数据的最佳描述. 引入核函数后,上述的目标优化问题变为

$$\max L = \sum_{i=1}^{n}\alpha_i K(x_i \cdot x_i) - \sum_{i=1,j=1}^{n}\alpha_i\alpha_j K(x_i \cdot x_j), \tag{5}$$

$$\text{s.t.} \sum_{i=1}^{n}\alpha_i = 1, 0 \leqslant \alpha_i \leqslant C. \tag{6}$$

只有少量的样本满足上式的等号,其对应的 α_i 不为 0,称之为支持向量. 利用任一支持向量可以求出超球的半径:

$$R^2 = K(x_k \cdot x_k) - 2\sum_{i=1}^{n}a_i K(x_i \cdot x_k) + \sum_{i=1,j=1}^{n}\alpha_i\alpha_j K(x_i \cdot x_j). \tag{7}$$

对于一个测试样本,可以依据下面的条件判断是否接受其为该类对象,如果满足

$$\| z - a \|^2 = (z \cdot z) - 2\sum_{i=1}^{n}\alpha_i K(z \cdot x_i) + \sum_{i=1,j=1}^{n}\alpha_i\alpha_j K(x_i \cdot x_j) \leqslant R^2,$$

则认为是该类样本,否则拒绝.

根据 Tax 和 Duin 的结论[9]:高斯型核函数相对于线性核函数和多项式核函数具有更好的性能. 其形式为 $K(x_i,x_j) = e^{\frac{\|x_i - x_j\|}{\sigma^2}}$. 本文采用的是高斯型核函数.

1.2　参数选择

由上可知,聚类的好坏取决于如何调节惩罚因子 C 和高斯核函数的宽度 σ. 惩罚因子 C 实现错分样本的比例和算法复杂度之间的折衷. 它的选取一般由具体的问题而定, C 越小,对经验误差的惩罚小,学习的复杂度小而经验风险值较大,超球的边界越平滑,得到的支持向量的个数越少. SVDD 性能的优劣还受到高斯核参数 σ 的影响,σ 越小,超球的边界越紧致,得到的支持向量的个数越多. 如何得到最佳的参数,目前还没有统一的方法,常用的方法是采用试凑法,针对某个特定的问题,通过多次尝试得到满意的结果,例如交叉验证法和网格搜索法[12]. 本文将聚类评估准则作为目标函数,采用智能算法进行参数寻优,从而避免反复凑试参数的繁琐和耗时,并且通过在酵母基因表达数据的聚类分析中进

行实验,验证了算法的有效性.

2 改进的支持向量数据描述算法

2.1 聚类有效性评价准则

评价聚类算法的有效性,即能否将基因正确的聚类,对于选择聚类算法来预测未知基因的功能有一定的指导意义. 目前,对于生物数据聚类评价使用最多的是由 Yeung 提出的聚类有效性的内部检验方法——FOM 法[13]. 假设 n 个基因被聚成 k 类:C_1,C_2,…,C_k,令 $\boldsymbol{R}(g;e)$ 表示原始数据矩阵中基因 g 在条件 $e(e=1,\cdots,m)$ 下的表达水平,$\mu_{C_i}(e)$ 表示 C_i 类内的基因在条件 e 下的平均表达水平,则

$$\mathrm{FOM}(k) = \sum_{e=1}^{m} \mathrm{FOM}(e,k).$$

式中:

$$\mathrm{FOM}(e,k) = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}(R(x,e)-\mu_{C_i}(e))^2}. \tag{8}$$

$\mathrm{FOM}(e,k)$ 为条件 e 下所有基因表达水平距离均差平方和的方根,反应类内的变异. $\mathrm{FOM}(k)$ 表示把整个数据集聚成 k 类时,所有类的类内变异总和. FOM 越小就表明类内的变异越小,算法的聚类效果越好.

由于 FOM 法并没有考虑使用类间的信息,本文将类间信息融入聚类有效性评价中,定义 $\mathrm{Dis}(k)$ 反映类间的变异:

$$\mathrm{Dis}(k) = \sum_{e=1}^{m} \mathrm{Dis}(e,k).$$

式中:

$$\mathrm{Dis}(e,k) = \sqrt{\frac{1}{n}\sum_{i,j=1}^{k}(\mu_{C_i}(e)-\mu_{C_j}(e))^2}. \tag{9}$$

式中:$\mu_{C_i}(e)$ 表示 C_i 类内的基因在条件 e 下的平均表达水平.

定义:

$$\mathrm{Val} = \mathrm{FOM}(k) - \mathrm{Dis}(k). \tag{10}$$

如果类内变异越小,类间变异越大,则 Val 值越小,聚类的质量越高. 本文将 Val 作为评价聚类算法的目标函数,用来引导 SVDD 的参数选取.

2.2 非目标样本的引入

为了提高算法收敛的速度,本文采用有监督的 SVDD 进行训练,即加入非目标样本类别信息,以保证超球边界两边都得到支撑,同时对于边界噪声也有一定的抑制作用.

加入非目标样本类别信息后的约束条件变为

$$y_i(R^2+\xi_i-\|x_i-a\|^2)\geqslant 0,\xi_i\geqslant 0,\forall i, \tag{11}$$

式中:$y_i=\begin{cases}1, & \text{目标样本;}\\ -1, & \text{非目标样本.}\end{cases}$

此时的 Lagrange 极值问题变为

$$\max L = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i y_i[R^2+\xi_i-((x_i\cdot x_i)-2ax_i+a\cdot a)] - \sum_{i=1}^{n}\gamma_i\xi_i,$$

式中:$\alpha_i\geqslant 0,\gamma_i\geqslant 0$.

求解该式的极值,分别对 R、a 和 ξ_i 求偏导并令其为 0,得到

$$\sum_{i=1}^{n}\alpha_i y_i = 1, a = \sum_{i=1}^{n}\alpha_i y_i x_i,$$
$$C-\alpha_i y_i-\gamma_i = 0,\forall i.$$

如果设 $\alpha_i'=\alpha_i y_i$,并用核函数代替内积运算,则会得到一个和式(5)形式一样的目标函数,

$$\max L = \sum_{i=1}^{n}\alpha_i' K(x_i\cdot x_i) - \sum_{i=1,j=1}^{n}\alpha_i'\alpha_j' K(x_i\cdot x_j). \tag{12}$$

可以看出,引入非目标样本,只是增加了训练样本的数量,并没有影响训练的复杂度.

2.3 参数的优化

模拟退火算法是一种全局最优算法,本文利用该方法来搜索 SVDD 的核函数参数 σ 和 C. 采用上述的聚类有效性评价准则作为适应值,通过模拟退火算法进行搜索迭代,找到满意的 SVDD 参数. 其具体步骤和算法如下:

1)初始化参数设置;

2)随机产生初始 SVDD 模型,计算其适应值 fitness = Val;

3)通过随机扰动产生新的 SVDD 模型,计算新的适应值 finess*;

4)按照 Metropolis 准则接受或放弃新的参数;

5)重复 3)和 4)完成一次 Metropolis 迭代过程;

6)判断适应值是否满足要求,如果满足,则算

法结束,输出 SVDD 模型;否则,按照一定的退火方案衰减 t,重复 3)、4)、5)继续寻优,直至得到满意的结果.

3　实验结果与分析

3.1　实验数据

本文采用已知类别的 446 条酵母(Yeast)细胞生长周期的表达数据作为实验数据,每个基因表达数据是 80 维的,分别表示不同的实验条件下、不同时间点的基因表达水平值,根据其细胞周期内的表达峰值分为 5 类:166 个 G_1 类型、115 个 S 类型、56 个 G_2 类型、42 个 M 类型、67 个 M/G_1 类型. 其中 G:生长期(growth);S:合成期(synthesis);M:分裂期(mitosis). 在不同阶段可能会有交界. 在聚类分析之前,先对基因表达数据进行了归一化处理.

3.2　结果与分析

对于多类分类问题,当训练某一类的最小超球时,非目标样本从其他各类中利用“自举法”产生,采用本文算法对实验数据进行聚类,仿真结果如下.

表 1 对比了网格搜索法和模拟退火法在 SVDD 参数寻优中的结果,与传统的网格搜索法相比,采用模拟退火算法来确定 SVDD 参数不仅缩短了训练时间,减少了支持向量的数目,提高了计算效率,而且一定程度上改善了训练样本和测试样本的聚类正确率.

表 1　网格搜索法和模拟退火方法的比较

Table 1　Comparison of cross-validation and SA

性能比较	网格搜索法	模拟退火法
(C,σ)最优值	(21.365,2.245)	(23.487,2.341)
训练时间/s	8.68	2.79
支持向量数/个	138	85
训练样本聚类正确率/%	93.96	94.14
测试样本聚类正确率/%	89.83	89.94

表 2 给出了不同聚类算法对测试数据集的聚类正确率,可见,SVDD 算法的聚类正确率相对于常用的聚类算法有了很大提高,验证了通过寻找半径最小的最优超球来覆盖属于同一类的数据样本点的 SVDD 算法可以得到更紧凑的聚类结果,不仅减少了误判率,同时,这种类似于认知样本的聚类算法,当新的类别出现时,不需要重新训练全部样本.

表 2　不同聚类算法正确率比较

Table 2　Accuracy comparison of different clustering methods

聚类算法	聚类正确率/%					
	G_1	G_2/M	M/G_1	S	S/G_2	平均值
Aagglo-merative	79.03	63.93	58.35	60.87	54.01	63.24
K-mean	84.21	76.92	68.42	72.73	70.82	74.62
FCM	90.70	88.89	76.92	69.23	66.67	78.48
SOM	97.22	93.33	65.00	64.71	80.00	82.80
SVDD	100	90.48	87.89	85.85	92.12	91.27

图 1 是各种聚类算法的 Val 值在不同类别数下变化的曲线. 由于 Val 融入了类间信息,对聚类的有效性评价更加客观和全面. 图 1 中的曲线高低和表 2 的结果基本一致,说明使用 Val 指导聚类算法的选择有一定的意义. 从图 1 中也可以看出,本文算法的 Val 值最小,这不仅由于 SVDD 算法本身的特点,而且在训练时引入非目标样本,使得超球更加紧凑,对于算法的收敛也有促进作用.

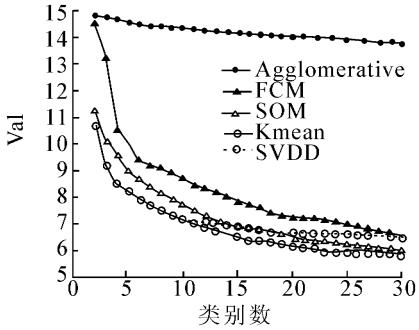


图 1　各聚类算法的 Val 值

Fig. 1　The Val of different clustering methods

4　结束语

聚类是基因表达数据分析的重要步骤,通过聚类分析,可以预测基因功能,构建基因调控网络,进而对内部的生命现象进行解释. 采用支持向量数据描述算法,通过寻找最优超球来覆盖样本数据,较好地解决了常用聚类算法存在的误判问题. 为了避免寻找核函数参数和惩罚因子的繁琐工作,改进了聚类有效性评估准则,以此作为目标函数,通过模拟退火优化算法得到 SVDD 的参数,在训练过程中加入非样本数据,从而提高了计算效率、聚类精度和对边界噪声的抑制能力. 将本文算法应用在酵母基因表达数据聚类分析中,结果验证了其有效性和快速性.

参考文献:

[1]BEISEN M, SPELLMAN P T, BROWN P O,et al. Cluster analysis and display of genome-wide expression patterns [J]. Proc Natl Acad Sci, 1998,95(12):14863-14868.

[2]P TAMAYO, SLONIM D, MESIROV J, et al. Interpreting patterns of gene expression with self organizing maps[J]. Proceedings of National Academy of Sciences,1999,96(6): 2907-2912.

[3]SUGIYAMA A, KOTANI M. Analysis of gene expression data by using self-organizing maps and k-means clustering [J]. Neural Network, 2002(5): 1342-1345.

[4]HERRERO J, VALENCIA A, DOPAZO J. A hierarchical unsupervised growing neural network for clustering gene expression patterns[J]. Bioinformatics, 2001:17(2):126-136.

[5]YUNG Y, RUZZO W. An empirical study on principal component analysis for clustering gene expression data [J]. Bioinformatics, 2001:17(9):763-774.

[6]BURGES C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery,1998,2(2):1-47.

[7]WU Fangxiang,ZHANG W J,KUSALIK A J. Dynamic model-based clustering for time-course gene expression data[J]. Journal of Bioinformatics and Computational Biology,2005,3 (4):821-836.

[8]JI X L, YUAN Y, LI Y D, SUN Z R. HMMGEP:clustering gene expression data using hidden Markov models [J]. Bioinformatics, 2004, 20(11):1799-1800.

[9]TAX D M J, DUIN R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999,20(11/13): 1191-1199.

[10]CERVANTES J, LI Xiaoou, YU Wen, LI Kang. Support vector machine classification for large data sets via minimum enclosing ball clustering[J]. Neurocomputing,2008, 71:611-619.

[11]边肇祺,张学工.模式识别[M].2版.北京:清华大学出版社,2002:296-304.

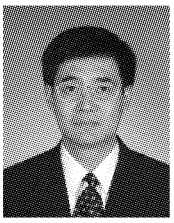
[12]BOSER B E, GUYON I M, VAPNIK V N. A training algorithm for optimal margin classifier[C]//Proc 5th Annu ACM Workshop on Compute Learning Theory. Haussier, Ed,1992:144-152.

[13]YEUNG K Y, HAYNOR D R, RUZZO W L. Validating clustering for gene expression data[J]. Bioinformatics, 2001,17(4):309-318.

[14]岳 峰, 孙 亮,王宽全,王永吉,左旺孟.基因表达数据的聚类分析研究进展[J]. 自动化学报,2008,34 (2):113-120.
YUE Feng, SUN Liang, WANG Kuanquan, WANG Yongji, ZUO Wangmeng. State-of-the art of cluster analysis of gene expression data[J]. Acta Automatica Sinica, 2008,34(2):113-120.

作者简介:

季瑞瑞,女,1978年生,讲师,博士研究生,主要研究方向为人工智能与模式识别、生物信息处理与建模.发表学术论文6篇,其中被EI检索3篇.



刘 丁,男,1957年生,教授,博士生导师,中国人工智能学会常务理事,主要研究方向为复杂系统的建模与控制、智能机器人、系统控制等.发表学术论文100余篇,其中被SCI、EI检索50余篇.