基于有限混合多变量 t 分布的鲁棒聚类算法*)

余成文　郭　雷

(西北工业大学自动化学院　西安 710072)

摘　要　在用混合模型聚类时，聚类数据中存在局外点是非常困难的问题。为了提高混合拟合的鲁棒性，本文用混合 t 模型替代混合高斯模型，来拟合含有背景噪音的多变量多高斯分布数据；提出了两个求解混合 t 模型的修改版期望最大化(EM)算法，并将它们与模型选择准则集成在一起，应用一个组合规则成分灭绝策略选择聚类成分数，得到两个对应的鲁棒聚类算法。对含有背景噪音的多个高斯成分进行不同聚类算法的大量实验表明，本文的鲁棒聚类算法能自动选择最佳的聚类成分数，相对于混合高斯模型的聚类方法，鲁棒性增强很多；相对于传统求解混合 t 模型(EM/ECM)的聚类方法，能有效避免其严重依赖初始值和易收敛至参数空间边界的缺点，具有较强的鲁棒性和较快的收敛速度。

关键词　局外点，鲁棒聚类，混合 t 模型，期望最大化算法，模型选择准则

Robust Clustering Based on Finite Mixtures t Distribution

YU Cheng-Wen　GUO Lei

(College of Automation, Northwestern Polytechnical University, Xi'an 710072)

Abstract　Providing protection against outlier in clustering data is a difficult problem for mixtures models fitting. In this paper, we consider the fitting of mixtures t distributions alternative to mixtures normal distributions for multi-component gauss data with background noise, to improve the robustness of fitting. We propose two modified versions of EM algorithm and integrate them with a model selection criterion respectively, then we get two robust clustering algorithms which can avoid the drawbacks of traditional algorithms (EM/ECM) for solving mixtures t models- highly dependent on initialization and may converge to the boundary of the parameter space, and can also select the number of clusters component automatically by a combined component annihilation strategy. Experiment results show the contrast among different algorithms and demonstrate the effectiveness of our algorithms.

Keywords　Outlier, Robust clustering, Mixtures t distribution, Expectation maximization, Model selection criterion

1　引言

有限混合模型作为一种数据的统计建模工具，已经在模式识别、计算机视觉和机器学习等领域获得了广泛的应用[1]。在聚类研究中引入有限混合模型，可以将单变量或多变量的数据用概率建模的方法来表达，这有利于用公式化的方式来处理聚类问题，如聚类成分数的选择，模型有效性的评估。混合高斯模型由于计算上的便利，在聚类方法研究中应用较多，但高斯分布由于尾巴较短，容易受噪音干扰；而混合 t 模型由于其 t 分布具有较长的尾巴，并由于其与鲁棒 M 估计器的联系，对非典型成员进行降低权值处理，因此，相对于混合高斯模型，可以获得较强的鲁棒性。

多变量 t 分布的参数最大似然(ML)估计一般用 EM 算法来求解。EM 算法迭代地寻找(通过 E 步)和最大化(通过 M 步)Q 函数来逼近参数的概率模型，L 函数-对数似然函数。对于单个 t 分布，Meng 和 Rubin[7]用一系列受限制的最大化 CM 步来替代 M 步，解决 M 步比较难求的问题，得到期望条件最大化算法(ECM)。Liu 和 Rubin[2]为了加快 ECM 算法收敛速度，对其进行两处修改，得到一个多周期版本的 ECM 算法和双期望条件最大化(ECME)算法，后来 Liu[3]又在此基础上提出了收敛速度更快的新版本 ECME 算法。Peel D 和 McLachlan G J[4]提出混合 t 模型，用标准 EM 求解 t 混合模型参数的 ML 估计，并给出了一个 ECM 算法的应用。Celeux 等[5]为了提高有限混合估计中 EM 算法的速度，提出了串行升级参数的混合模型成分方式 EM 算法(CEM^2)。然而，传统 EM 及其扩展的算法有几个缺点：高度依赖初始化，容易收敛至参数空间的边界，为了解决这个问题，Figueiredo 和 Jian A K[6]提出一个最小信息长度(MML)收敛准则，通过使用了一个集成估计和模型选择准则的策略，成功应用于混合高斯模型。

结合多周期版本 ECM 和 CEM^2 算法的思想，本文提出了两个求解混合 t 模型的修改版 EM 算法，将这两个算法与 MML 准则相结合，并应用一个组合规则成分消灭策略，得到两个鲁棒聚类算法。通过实验表明，本文的鲁棒聚类算法能自动选择成分数、能有效避免传统 EM 算法的缺点并具有较强的鲁棒性，实验给出了多种算法的性能对比。

2　学习有限混合 t 模型

2.1　有限混合多变量 t 分布

设 $y_1 \cdots y_n$ 为 p 维随机变量 Y 的 n 个样本值，用基于混合 t

*)基金项目：国家自然科学基金项目(60175001)资助。余成文　博士生，主要从事计算机视觉与模式识别研究；郭　雷　教授，博导，主要研究方向：神经计算，模式识别等。

模型的方法来建模这些数据，Y 的概率密度函数可以表示为：

$$f(y|\Psi)=\sum_{i=1}^{m}\pi_i f(y|\mu_i,\sum_i,v_i) \tag{1}$$

式中，$\pi_1,\cdots,\pi_m$ 为混合比例，每个值均非负并且所有值的和为 1，$f(y|\mu_i,\sum_i,v_i)$ 表示 p 维变量 t 分布的概率密度函数，μ_i 为位置参数，$\sum_i$ 为正定内积矩阵，v_i 为自由度参数。Ψ 为完整参数集，$\Psi=(\pi_1,\cdots,\pi_m,v_1,\cdots,v_m,\theta_1,\cdots,\theta_m)$，$\theta_i$ 由 μ_i 和 $\sum_i$ 组成。

2.2 最大似然估计与 EM 算法

混合密度估计问题是一个典型的丢失数据问题，用 EM 算法比较有效。在 EM 算法的框架中，具有混合概率密度的随机变量由于单个观测样本具体属于哪个成分未知，经常被视为不完整数据。对于具有 t 混合概率密度的随机变量 Y，完整数据为 $X=\{Y,Z,U\}$，其中 $Z=(z_1,\cdots,z_n)$ 为二进制数据，与样本数据相关，用来指示哪个成分产生该样本，如 z_1，$=(z_1^{\{1\}},\cdots,z_1^{\{n\}})$，$z_1^{\{i\}}=1$，$z_1^{\{j\}}=0(i\neq j)$，表示第一个样本由第 i 个成分产生；$U=(u_1,\cdots,u_n)$ 为针对 t 分布特点的附加的丢失数据[4]。

EM 算法迭代地应用两步产生一系列的估计 $\{\hat{\Psi}^{(k)},k=0,1,2,\cdots\}$，直到按某个准则收敛。

E 步：给定 Y 和当前的估计 $\hat{\Psi}^{(k)}$，计算完整对数似然 $\log(p(X|\Psi))$ 的条件期望。由 $\log(p(X|\Psi))$ 与丢失的数据 Z 和 U 具有线性关系，可以只求解条件期望 $T=E(Z|Y,\Psi)$，$W=E(U|Y,Z,\Psi)$ 然后将其代入 $\log(p|X|\Psi))$，得到 Q 函数 $Q(\Psi;\hat{\Psi}^{(k)})$。第 i 个成分使用当前的拟合 $\hat{\Psi}^{(k)}$，升级先验 $\pi_i^{(k)}$ 确定在 y_j 属于混合模型中第 i 个成分后的后验

$$\tau_{ij}^{(k)}=\frac{\pi_i^{(k)}f(y_j;\Psi_i^{(k)})}{\sum_{i=1}^{m}\pi_i^{(k)}f(y_j;\Psi_i^{(k)})} \tag{2}$$

升级当前拟合 $\hat{\Psi}^{(k)}$ 的鲁棒调节参数

$$u_{ij}^{(k)}=\frac{v_i^{(k)}+p}{v_i^{(k)}+\delta(y_i;\mu_i^{(k)},\sum_i^{(k)})} \tag{3}$$

M 步：根据 $\hat{\Psi}^{(k+1)}=\arg\max_{\Psi}Q(\Psi;\hat{\Psi}^{(k)})$ 升级参数

$$\pi_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k)}/n \quad (i=1,\cdots,k) \tag{4}$$

$$\mu_i^{(k+1)}=\sum_{j=1}^{n}\tau_{ij}^{(k)}u_{ij}^{(k)}y_j/\sum_{j=1}^{n}\tau_{ij}^{(k)}u_{ij}^{(k)} \tag{5}$$

$$\sum_i^{(k+1)}=\frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}u_{ij}^{(k)}(y_j-\mu_i^{(k+1)})(y_j-\mu_i^{(k+1)})^T}{\sum_{j=1}^{n}\tau_{ij}^{(k)}u_{ij}^{(k)}} \tag{6}$$

在升级 $v_i^{(k)}$ 时没有闭式解，$v_i^{(k+1)}$ 通过如下非线性方程来求解

$$\{-\Psi(0.5v_i)+\log(0.5v_i)+1+(1/\sum_{j=1}^{n}\tau_{ij}^{(k)})\sum_{j=1}^{n}\tau_{ij}^{(k)}(\log u_{ij}^{(k)}-u_{ij}^{(k)})\}+\Psi(0.5(v_i^{(k)}+p))-\log(0.5(v_i^{(k)}+p))\}=0 \tag{7}$$

式中，$\Psi(x)=(\partial\Gamma(x)/\partial x)/\Gamma(x)$。

本文在用信赖域方法(Trust Region)求解(7)式时，有时会出现溢出现象($v_i^{(k+1)}<0$，出现极少)，当出现这种情况，$v_i^{(k)}$ 不进行升级，$v_i^{(k+1)}=v_i^{(k)}$。

3 鲁棒聚类算法

3.1 提出的新版本 EM 算法

算法 1 成分方式多周期 ECM(CMECM)

Guass-Seidel 策略是一个将优化问题分解为一系列同等方式的最大化问题的流程[8]，Celeux 等[5]将流行的最接近方法[9]和 Guass-Seidel 策略融合在一起，得到混合模型成分方式 EM 算法(CEM²)，CEM² 算法以串行方式替代并行方式搜索参数空间，来防止陷入局部最小困境。CEM² 每次只升级一个成分，其它参数保持不变，升级成分的顺序可以是随机的，指定的或者自适应变化的。

ECM 算法[7]是用来解决含未知自由度 v 单成分 t 分布的 ML 估计中求解 M 步困难问题，被 Peel D 和 McLachlan G J[4]应用在混合 t 分布中。ECM 算法将参数空间划分为几个部分，并用相同数量的 CM 步替代 M 步。多周期版本的 ECM 算法[2]在每两个 CM 步间有个附加的 E 步，收敛比 ECM 快。

结合 CEM² 算法和多周期版本的 ECM 算法的思想，得到一个拟合混合 t 分布的新版本 EM，CMECM。将 Ψ 划分为 (Ψ_1,Ψ_2)，其中，$\Psi_1=(\pi_1,\cdots,\pi_m,\theta_1,\cdots,\theta_m)$，$\Psi_2=(v,\cdots,v_m)$。为了简化表示，成分相继升级，$i=1,\cdots,m$，在 m 次迭代后重复进行，因此，在第 k 次迭代的成分为 $i=k-\lfloor k/m\rfloor m+1$，其中 $\lfloor\cdot\rfloor$ 表示取整。CMECM 算法的第 k 次迭代如下：

E 步：对 $j=1,\cdots,n$，用(2)式计算 $\tau_{ij}^{(k)}$，用(3)式计算 $u_{ij}^{(k)}$

CM 步 1：Ψ_2 固定在 $\Psi_2^{(k)}$，通过最大化 $Q(\Psi;\hat{\Psi}^{(k)})$ 计算 $\Psi_1^{(k+1)}$。即，v_i 固定在 $v_i^{(k)}$，分别用(4)，(5)，(6)式升级 $\pi_i^{(k+1)}$，$\mu_i^{(k+1)}$，$\sum_i^{(k+1)}$；当 $\ell\neq i$，$\pi_l^{(k+1)}=\pi_l^{(k)}$，$\theta_l^{(k+1)}=\theta_l^{(k)}$。

附加 E 步：用 $(\Psi_1^{(k+1)},\Psi_2^{(k)})$ 替代 $(\Psi_1^{(k)},\Psi_2^{(k)})$，对 $j=1,\cdots,n$，用(3)式计算 $u_{ij}^{(k)}$。

CM 步 2：Ψ_1 固定在 $\Psi_1^{(k+1)}$，通过最大化 $Q(\Psi;\hat{\Psi}^{(k)})$ 计算 $\Psi_2^{(k+1)}$。即，用(7)式升级 $v_i^{(k+1)}$。

算法 2 扩展形式 CMECM 算法(ECM²)

注意到求解非线性方程(7)比较费时，并且 CMECM 算法中的附加 E 步是个不完全的 E 步，仅用到 $u_{ij}^{(k)}$ 的信息。因此，可将 CEM² 算法仅用于升级 Ψ_1，在 m 次迭代后得到 $\tau_{ij}^{(k)}$ 的最新信息，然后与 $u_{ij}^{(k)}$ 一起升级 $v_i^{(k)}(i=1,\cdots,m)$。这样，得到 CMECM 算法的扩展，ECM²。

3.2 最小信息长度(MML)收敛准则

Figueiredo 等[6]派生出一个模型参数选择准则—MML 准则，并应用于混合高斯模型。本文将 MML 修改使其适应混合 p 变量 t 分布，得到 $\hat{\Psi}=\arg\min_{\Psi}\{\mathcal{L}(\Psi,Y)\}$，其中，

$$\mathcal{L}(\Psi,Y)=(N/2)\sum_{i=1}^{m}\log(\pi_i)-\log p(Y|\Psi)+[m(N+1)/2][1+\log(n/12)] \tag{8}$$

式中，$N=1+p+p(p+1)/2$ 表示具体到每个成分的参数个数。通过 *EM* 算法最小化(8)式的 M 步有如下形式，

$$\pi_i^{(k+1)}=\frac{\max\{0,\sum_{j=1}^{n}\tau_{ij}^{(k)})-N/2\}}{\sum_{i=1}^{m}\max\{\{0,(\sum_{j=1}^{n}\tau_{ij}^{(k)})-N/2\}} \tag{9}$$

当 $\pi_i^{(k+1)}>0$，其它参数按原来的 EM 升级。这样 MML 和 EM 集成在一起：对于固定的成分数(当 $\pi_i^{(k+1)}=0$，第 i 个成分被消灭)，聚类数据的模型拟合过程由 EM 表示，对聚类数据最好的模型表示由 MML 来选择。

3.3 成分数目的估计

混合模型中成分数的估计是混合拟合成功的关键也是难点。常用确定成分数的方法有确定性方法和随机重采样方法。随机重采样方法由于计算量大，比较费时。本文用线性确定退火方法来确定成分数，思想是，设定一个包含真实成分数 i_{true} 的有限范围 $i_{min}\sim i_{max}$，从最大值 i_{max} 开始向 i_{min} 方向搜索，期间用 EM 算法来计算各成分的参数，最后用 MML 准则选择聚类数据最佳的成分数。值得注意到是，由于(9)式也带有消灭成分的性质，实质上消灭成分是由确定性规则和(9)式中收敛规则共同完成的，因此可以引入一个新的概念－组合规则成分灭绝。组合规则成分灭绝由强规则成分灭绝和强制性成分灭绝组成，强规则成分灭绝来源于 MML 准则，由(9)

式中 $N/2$ 决定，强制性成分灭绝来源于 DA，由 MML 收敛准则中的收敛域值 γ 决定。

3.4 鲁棒聚类算法(RCAs)

根据 3.1，3.2，3.3 节的描述，得到对应于两个新版本 EM 算法(CMECM 和 ECM^2)的两个鲁棒聚类算法。在该两个鲁棒聚类算法中，新版本 EM 算法中的 $\pi_i^{(k+1)}$ 的升级用(9)式替代，在新版本 EM 算法收敛前(对 $\mathcal{L}(\Psi,Y)$，收敛表示为 $|\mathcal{L}^{(k-1)}-\mathcal{L}^{(k)}|<\gamma\mathcal{L}^{(k-1)}$)，成分灭绝由强规则成分灭绝控制，消灭混合比例 $\pi_i^{(k+1)}=0$ 的成分；在新版本 EM 算法收敛后，成分灭绝由强制性成分灭绝控制，强制消灭具有最小混合比例 $\pi_i^{(k+1)}$ 的成分，然后返回新版本 EM 算法，该流程重复进行，直到成分数目为 $i_{\min}$。对应具有最小 $\mathcal{L}(\Psi,Y)$ 的成分数目和模型参数被选择作为聚类数据的最佳表示。

4 实验

4.1 实验数据

实验数据由随机混合高斯分布和背景噪音两部分组成，随机混合高斯分布由 3 个均值和协方差已知的随机数据点按一定比例混合而成，背景噪音由均匀分布获得。

高斯混合分布参数为：混合比例，$\pi_1=\pi_2=\pi_3=1/3$；均值，$\mu_1=[0,0]^T$，$\mu_2=[0,4]^T$，$\mu_3=[0,-4]^T$；方差，$\sum_1=\sum_2=\sum_3=[3,0.2;0.2,0.4]$。背景噪音服从均匀分布，每个变量均在(−8,8)范围取值。高斯分布和背景噪音共 1000 个点，噪音比例为 0～20%。

4.2 实验结果与分析

验证鲁棒聚类算法的主要指标是：对噪音的鲁棒性和收敛速度。因此，本文比较多种算法：基于混合高斯模型(CEM^2)的聚类算法(CA)；基于混合 t 模型(标准 EM)的鲁棒聚类算法 1(RCA1)；基于混合 t 模型(CEM^2)的鲁棒聚类算法 2(RCA2)；基于混合 t 模型(提议的 CMECM)的鲁棒聚类算法 3(RCA3)；基于混合 t 模型(提议的 ECM^2)的鲁棒聚类算法 4(RCA4)。RCA1 使用贝叶斯推理准则(BIC)选择聚类成分数，其它算法用组合规则成分灭绝策略选择成分数。

对含噪音比例为 5%的样本数据，如图 1(a)，RCA3 算法运行如下，

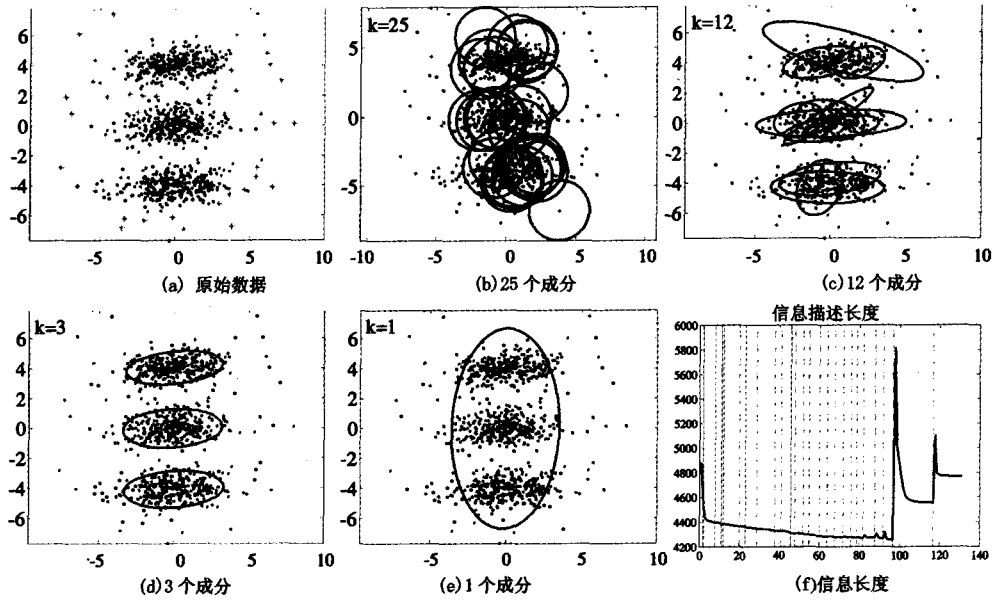

图 1 基于混合 t 模型的鲁棒聚类

初始随机选择的成分($i_{\max}=25$)如图 1(b)所示，RCA3 使用组合规则成分灭绝策略直到最小成分数($i_{\min}=1$)，如图 1(e)。信息长度的变化如图 1(f)所示，图中垂直实线表示强规则成分灭绝，垂直虚线表示强制性成分灭绝。图 1(f)给出了当成分数为 3 时(对应图 1(d))得到的最小信息长度。对上述算法进行大量试验，得到如图 2 的统计结果。

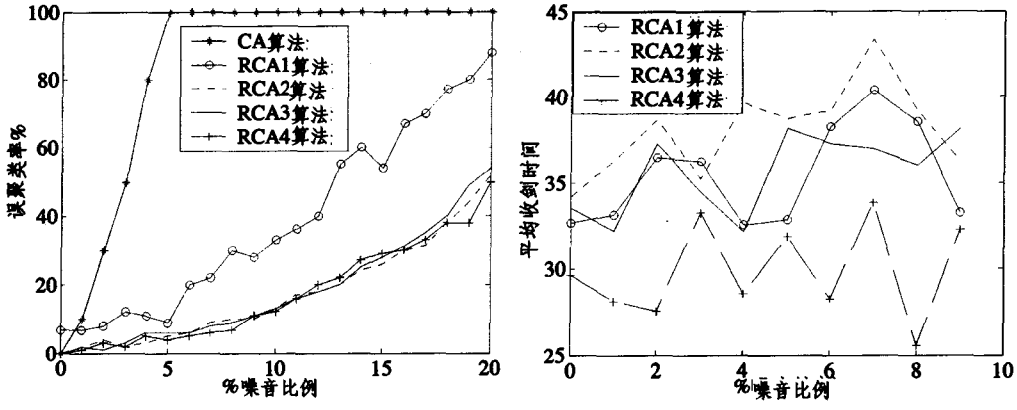

图 2 算法性能对比

图2(a)的结果表明,对含有背景噪音(一定范围)的混合高斯数据,用混合 t 模型替代混合高斯模型,鲁棒性明显增强;在基于混合 t 模型的鲁棒聚类算法中,用组合规则成分灭绝策略选择成分数的聚类算法比用 BIC 准则选择成分数的聚类算法鲁棒。值得注意的是,用组合规则成分灭绝策略选择成分数的聚类算法在噪音为0时,误聚类率接近0。图2(b)的结果表明,使用本文提议的 CMECM 和 ECM^2 的鲁棒聚类算法(RCA3 和 RCA4)比使用 CEM^2 的鲁棒聚类算法 RCA2 收敛快,RCA3 与 RCA1 收敛速度相当,RCA4 与 RCA1 收敛速度快。

结论 结合多周期版本 ECM 算法和 CEM^2 算法的思想,本文提出了两个求解混合 t 模型的修改版 EM 算法,通过集成 MML 收敛准则,使用组合规则成分灭绝策略,得到两个鲁棒聚类算法。实验表明,本文的鲁棒聚类算法结合了混合多变量 t 分布,多周期版本 ECM,CEM^2 算法,MML 准则和 DA 的优势,与混合高斯模型相比,对局外点的鲁棒效果强很多,与传统混合 t 模型求解方法 EM/ECM 相比,对局外点要鲁棒并且收敛速度更快。

参 考 文 献

1 McLachlan G, Peel D. Finite Mixture Models. New York: John Wiley & Sons, 2000

2 Liu C, Rubin D B. ML Estimation of the t Distribution using EM and its Extensions, ECM and ECME, Statistica Sinica, 1995(5): 19~39

3 Liu C. ML Estimation of the Multivariate t Distribution and the EM Algorithm. Journal of Multivariate Analysis, 1997, 63: 296~312

4 Peel D, McLachlan G J. Robust Mixture Modeling using the t Distribution, Statistics and Computing, 2000(10): 339~348

5 Celeux G, Chretien S, Forbes F, Mkhadri A. A Component Wise EM Algorithm for Mixtures. Journal of Computational and Graphical Statistics, 2001, 16(10): 697~712

6 Figueiredo MAT, Jain AK. Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3):381~396

7 Meng XL, Rubin DB. Maximum Likelihood Estimation via the ECM Algorithm: a general Framework. Biometrika, 1993, 80: 267~278

8 Ciarlet PG. Introduction to Numerical Linear Algebra and Optimization. Cambridge Texts in Applied Mathematics: Cambridge University Press, 1988

9 Chretien S, Hero AO. Generalized Proximal Point Algorithms and Bundle Implementation: [Technical Report. CSPL 313]. The University of Michigan, Ann Arbor, USA, 1998

(上接第 182 页)

每次循环的时间复杂度为 $O(l^2)$,所以,搜索一次的总时间复杂度为 $O(l^3)$。通常 $m>1$,并且搜索 t 次 SVr 才可能恒定,精确搜索才能终止,时间复杂度将为 $O(t/m\times l^3)$。由于 $\frac{t}{m}\ll l$,因此在 $O(l^3)$ 时间范围内一般能解。总的空间复杂度为 $2\times l\times(n\times 1)+3\times k\times(n+1)+3(k+1)\times(k+1)+k\times k+(k+1)\times 2$。$k$ 为训练子集样本数,$(n+1)$为样本维数。

不精确搜索的最好情形是只执行一次搜索就满足不精确搜索的终止条件,总时间复杂度为 $O(l^2)$。不精确搜索的最坏情形等同于精确搜索的最坏情形。通常情况下,不精确搜索仅需要搜索几次就能满足终止条件,总时间复杂度通常为 $O(l^2)$。总的空间复杂度为 $2\times l\times(n+1)+3\times k\times(n+1)+2\times(k+1)\times(k+1)+k\times k+(k+1)\times 2$。

2.3.3 *与 LS-SVM 线性方程组解法和直接的分解算法的比较*

从(7)、(8)式得知,LS-SVM 线性方程组解法的线性方程组系数矩阵为实对称矩阵。求该线性方程组系数矩阵的时间复杂度为 $O(l^2)$;解此线性方程组的时间复杂度为 $O((l+1)^3)$;总的空间复杂度为 $2\times l\times(n+1)+2\times(l+1)\times(l+1)+(l+1)\times 2$。使用从近似平面出发逐步搜索支持向量,构建 LS-SVM,核矩阵的规模从 $l\times l$ 变为了 $k\times k$。由于 $k\ll l$,最大耗时步骤由解线性方程组转变为构造新的训练样本子集。最坏情形下,两种搜索的时间复杂度为 $O(l^3)<O((l+1)^3)$;通常情况下,精确搜索在 $O(l^3)$ 能解,不精确搜索的时间复杂度为 $O(l^2)$;总的空间复杂度也下降为 $2\times l\times(n+1)+3\times k\times(n+1)+2\times(k+1)\times(k+1)+k\times k+(k+1)\times 2$。

直接的分解算法是将 l 个训练样本随机分为只包含 k 个训练样本的样本组,逐组求 SVM,然后将每组的 SV 组合在一起,训练最终的 SVM。如果求每组的 SVM 仍然采用 LS-SVM 线性方程组解法,其总时间复杂度为 $O((l/k+1)\cdot(k+1)^3)$,总的空间复杂度仍然为 $2\times l\times(n+1)+3\times k\times(n+1)+2\times(k+1)\times(k+1)+k\times k+(k+1)\times 2$。使用从近似平面出发逐步搜索支持向量构建 LS-SVM 算法的时间复杂度大于分解算法的时间复杂度,空间复杂度与分解算法的空间复杂度相一致。但是,该算法进行精确搜索建立的 LS-SVM,能保证收敛到 l 个训练样本直接解线性方程组建立的 LS-SVM,而分解算法不能。

结论 训练样本集的 SVM 存在近似超平面。根据 SV 分布于 SVM 超平面附近,也必然分布于其近似超平面附近的特点,可以通过向量距近似超平面的距离设计 SV 的逐步搜索算法。

从近似超平面到 SVR 算法实例——从多元回归平面构建 LS-SVM 的算法得证,从近似超平面出发逐步搜索 SV,建立 SVR 的算法能够降低计算时间复杂度,并且显著降低计算的空间复杂度。

从近似超平面到 SVR 算法进行精确搜索时,其结果可以收敛到 l 个训练样本直接建立的 SVR。

基于上述结果,根据出发搜索 SV 近似超平面和搜索过程中解 SVM 算法的不同,能够设计一类从近似超平面到 SVR 算法,用于解决回归问题,求取 SVR。

参 考 文 献

1 Avidan S. Subset selection for efficient SVM tracking [C]. Computer Vision and Pattern Recognition. In: Proc. 2003 IEEE Computer Society Conf., 2003, 1(1):85~92

2 Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers [C]. In D. Haussler, ed. Proceedings of the 5th annual ACM Workshop on computation learning theory, ACM Press, 1992. 144~152

3 Bradford C. The quickhull algorithm for convex hulls [C]. ACM Trans. Mathematical Software, 1996, 22(44): 469~483

4 Valyon J, Horvàth G. A Sparse Least Squares Support Vector Machine Classifier [C]. In: Proc. 2004 IEEE Neural Networks Conf., pp543~548

5 Vapnik V N. Support vector method [J]. Lecture Notes in Computer Science, 1997, 1327:263