Journal of Computer Applications
计算机应用,2014,34(8):2285 -2290
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2285-06
doi:10.11772/j.issn.1001-9081.2014.08.2285

多模型数据集的免疫鲁棒回归分析

徐雪松[1*],舒　俭[2]
(1.华东交通大学 电气与电子工程学院,南昌 330013;　2.华东交通大学 教务处,南昌 330013)
(*通信作者电子邮箱 cedarxu@163.com)

摘　要:针对传统多模型数据集回归分析方法计算时间长、模型识别准确率低的问题,提出了一种新的启发式鲁棒回归分析方法。该方法模拟免疫系统聚类学习的原理,采用B细胞网络作为数据集的分类和存储工具,通过判断数据对模型的符合度进行分类,提高了数据分类的准确性,将模型集抽取过程分解成"聚类""回归""再聚类"的反复尝试过程,利用并行启发式搜索逼近模型集的解。仿真结果表明,所提方法回归分析时间明显少于传统算法,模型识别准确率明显高于传统算法。根据8模型数据集分析结果,传统算法中,效果最好的是基于RANSAC的逐次提取算法,其平均模型识别准确率为90.37%,需53.3947 s;计算时间小于0.5 s的传统算法,其准确率不足1%;所提算法仅需0.5094 s,其准确率达到了98.25%。

关键词:多模型;模型集;鲁棒回归;免疫原理;启发式算法
中图分类号:TP301.6;TP18　　**文献标志码**:A

Immune robust regression analysis for data set of multiple models

XU Xuesong[1*], SHU Jian[2]
(1. *School of Electrical and Electronic Engineering, East China Jiaotong University, Nanchang Jiangxi* 330013, *China*;
2. *Academic Affairs Office, East China Jiaotong University, Nanchang Jiangxi* 330013, *China*)

Abstract: Classical regression algorithms for data set analysis of multiple models have the defects of long calculating time and low detecting accuracy of models. Therefore, a heuristic robust regression analysis method was proposed. This method mimicked the clustering principle of immune system. The B cell network was taken as classifier of data set and memory of model set. Conformity between data and model was used as the classification criteria, which improved the accuracy of the data classification. The extraction process of model set was divided into a parallel iterative trial including clustering, regressing and clustering again, by which the solution of model set was gradually approximated to. The simulation results show that the proposed algorithm needs obviously less calculating time and it has higher detecting accuracy of models than classical ones. According to the results of the eight-model data set analysis in this paper, among the classical algorithms, the best algorithm is the successive extraction algorithm based on Random Sample Consensus (RANSAC). Its mean model detecting accuracy is 90.37% and the calculating time is 53.3947 s. The detecting accuracy of those classical algorithms which calculating time is below 0.5 s is bellow 1%. By the contrary, the proposed algorithm needs only 0.5094 s and its detecting accuracy is 98.25%.

Key words: multiple model; model set; robust regression; immune principle; heuristic algorithm

0　引言

常规的回归分析方法,如最小二乘法,主要目的是寻找一组模型参数,使其按某种性能测度最佳拟合数据集中的所有数据。然而,由于数据集中难免会混入异常数据,拟合这些数据有时会对回归结果产生巨大影响,导致参数估计发生较大偏离。鲁棒回归分析研究的问题就是如何排除或减小这种异常数据对估计结果的影响。

一种直观的方法是在性能指标函数里给每个数据附加一个权值,让离差较大的数据权值较小,以减少其对估计结果的影响。这就是著名的M估计的基本原理[1]。M估计算法简单,计算速度快;缺点是崩溃点较低,即所能容忍的异常数据(离群点)的数目比例不高。为了提高算法的崩溃点,陆续出现了的LMS(Least Median Square)[2-3]、LTS(Least Trimmed Square)[4-5]、RANSAC(Random Sample Consensus)[6-7]等方法。LMS的做法是先随机选择数据点进行参数估计,然后计算所有数据在此模型下的离差,选择中位数作为评判标准。通过足够次数的抽样,选择中位数最小的模型参数作为最终结果。LTS与LMS类似,不同的是选择离差较小的前一半数据离差平方和最小作为评判标准。LTS与LMS的崩溃点较高,可以达到50%。其缺点是采用抽样试凑评判的方法,在大数据集的条件下,需要进行较多次数的抽样才能保证满意的结果,因而计算时间较长。如何在保证算法效果的同时,有效减小抽样次数,以缩短计算时间是人们重点关注的问题之一[8]。RANSAC也是一种通过抽样评判的方式获得模型参数的方法,RANSAC具有很高的崩溃点,广泛应用于图像数据处

收稿日期:2014-02-21;**修回日期**:2014-03-26。　**基金项目**:国家自然科学基金资助项目(61165005,51167005)。
作者简介:徐雪松(1970-),男,江西德兴人,副教授,博士,主要研究方向:多模型建模及控制;　舒俭(1971-),男,江西南昌人,高级实验师,硕士,主要研究方向:计算机系统及仿真。

理领域;缺点是计算时间长,不适合用于高维、大数据集的处理。

以上方法多侧重于单模型参数估计。现实生活中,有时多个模型的数据混杂在一起,需要从中估计出每个模型参数。例如:根据图像数据提取多条曲线的参数;根据采样数据估计多模型不确定系统的模型集等。在这些问题中,数据集中的异常数据以及来自于其他模型的正常数据都会对特定模型的回归分析产生较大的干扰。如何根据受污染的多模型数据集准确估计出所有模型的参数是一个较为困难的问题,也是当前鲁棒回归算法研究的重要内容之一[9-10]。

目前,多模型数据集的参数估计方法一般有两类:

1)逐次提取[11]。这类方法以单模型鲁棒回归算法为核心,先提取数据量大的主要模型,然后剔除相关数据,进行下一个模型的提取,直至所有数据被处理完毕。这类方法的优点是算法简单,但是要求不同模型在数据量上有主次之分;否则当离群数据远比有效数据多的时候,低崩溃点的鲁棒回归算法效果不好。

2)"聚类+回归"分析[12]。这一类方法先通过聚类将数据集分成不同的子集,然后对各子集分别采用单模型鲁棒回归算法进行参数估计。这类方法的关键在于找到合适的聚类工具,将数据集按模型的不同准确分类,错误的分类可能会严重影响参数估计效果。实际上,上述两种方法都试图一次性剥离各模型数据。然而,一次性从含大量离群点的数据集中获得准确的模型参数,或者一次性准确地按模型不同将数据集分类,都存在较大困难,因此两种方法的分析效果都有待于进一步提高。

生物免疫系统中针对不同的抗原能够进化出不同抗体,抗体集的学习过程本质上类似多模型数据集的参数估计过程。免疫系统利用 B 细胞网络作为数据集的分类与模型的存储工具,既不尝试一次性将数据准确分类,也不尝试一次性准确地识别某一种模型参数。而是通过一种"尝试聚类-模型抽取-修正聚类"的反复修正过程,实现模型集的并行进化和储存。本文借鉴上述工作机制,将免疫聚类与常规单模型鲁棒回归分析相结合,提出一种启发式多模型数据集鲁棒回归分析方法。仿真结果表明,该方法在具有大量离群点的多结构数据条件下,可以实现快速准确的模型集参数估计。

1 免疫系统与免疫聚类

生物免疫系统的主要功能在于识别并消灭入侵肌体的抗原性异物,保持肌体健康。一般分成非特异性免疫与特异性免疫。特异性免疫又称获得性免疫,是免疫系统针对特定抗原,通过后天学习,产生抗体,获得抗感染能力的过程,这是免疫系统适应外界能力的主要体现。

免疫系统识别抗原产生抗体的功能主要通过免疫 B 细胞实现。B 细胞来源于骨髓的多能干细胞,经过分化后形成成熟的 B 细胞进入脾脏、淋巴结,主要分布于脾小结、脾索及淋巴小结、淋巴索及消化道粘膜下的淋巴小结中。B 细胞受抗原刺激后,分化增殖为浆细胞,合成抗体,发挥体液免疫的功能。

B 细胞表面受体的可变区与抗原可以看成一种钥匙与锁的关系。可变区的分子排列多种多样,当它的分子排列与抗原分子排列相近时,抗原会刺激这一类 B 细胞产生免疫反应,开始克隆选择过程。在这个过程中,B 细胞通过不断的分裂增殖,并在增值过程中发生高频变异,其中与抗原结构更接近的细胞得到进一步的强化,从而筛选出与抗原高度亲和的细胞形成浆细胞分泌抗体。没有受到激励的 B 细胞将自动凋亡。免疫系统会定期补充新的 B 细胞以实现对抗原结构空间的遍历。此外,B 细胞存在独特型,这些独特型的功能类似抗原,能够被其他 B 细胞识别并抑制。所有 B 细胞以相互激励与制约的网络化形式存在。

免疫聚类是借鉴免疫系统抗体的学习和存储原理发展起来的一类智能学习方法[13]。由于该方法在具有良好的性能,目前已经得到了广泛的应用。一般来说免疫聚类有如下特征:

1)以聚类数据为抗原,聚类中心为 B 细胞,所有 B 细胞构成 B 细胞网络。

2)以聚类数据与 B 细胞的符合度为抗原细胞间的亲和度,抗原能够激励 B 细胞生存和进化;以模型相似度为 B 细胞之间的亲和度,B 细胞能抑制相似细胞的生存以减小网络冗余。

3)采用克隆选择的方法实现最优解集的搜索和进化,通过抗原与 B 细胞及 B 细胞之间的激励与抑制关系实现 B 细胞网络的更新。

4)采用群体进化的算法结构。

2 免疫鲁棒回归方法

2.1 基本思路

在多模型数据集回归分析时,不同模型间的数据相互干扰是主要问题。在提取特定模型的时候,其他模型的数据也是离群数据,这些数据的数量一般远大于单个模型数据量,因而要求回归算法具有很高的崩溃点,目前许多算法达不到这个要求。通过聚类将数据集分成若干个子集,让整体不占优的特定模型数据在数据子集上占优,然后即使采用常规的鲁棒回归方法也能准确地提取模型参数,这是一个较好的思路。问题是怎样聚类才能够将数据集按模型不同准确分类。目前"聚类+回归"分析想通过一次性聚类分析就将不同数据准确分开,比较困难。因为数据集通常在数据空间上并不一定有聚集特性,常规基于数据距离的聚类方式效果不一定好。

对于由模型集产生的数据,虽然按数据间距分,聚类特征并不明显,但是如果按模型符合度来分,即按照与模型的离差大小来分类,则数据具有明显的聚集特征。检验模型与实际模型的近似度越高,所拥有的高吻合度数据就越多,从这个角度看模型本身是一个良好的数据分类器。问题是这个分类器也是本文需要求解结果。

在免疫系统中,特异性 B 细胞靠能够识别更多的高亲和度抗原,获得更大的激励而胜出的,其机制类似基于模型的数据分类。在获得抗体之前,免疫系统也不知道抗原结构。免疫系统采取的方法是随机检测,然后进化优选。免疫系统中具有大量的不同结构的 B 细胞,这些 B 细胞相当于备选模型。它们根据受激励大小决定是否被激活。一旦激励大到一定程度,则开始克隆增生并进化优选,直至获得高亲和度抗体。类比到本文问题,也就是先通过随机测试寻找模型,然后按模型占有数据多少确定模型优劣,对优势模型进行进化优选,使得该模型更有优势,通过这种循环迭代完成模型集抽

取。由于抽取过程被分解成“尝试聚类-回归检验-优选后再聚类-再回归检验”的逐步逼近的过程。在模型逐步修正的同时,数据分类也越来越准确,而更准确的分类将导致更良好的回归效果,从而形成良性循环,所以不再要求一次性将好的模型挑出来,也不必要一次就将数据集分类好,因而能够有效降低对回归算法和聚类算法的要求。

考虑到免疫聚类算法集中体现了免疫系统这种抗体集的进化学习方式,本文以免疫聚类算法为基本框架,将它与常规单模型鲁棒回归分析方法相结合,给出了一种适合用于多模型数据集分析的启发式鲁棒回归方法,具体算法描述见 2.2 节。

2.2　算法描述

定义 1　B 细胞 i 为一组模型回归参数:

$$\boldsymbol{b}_i = \{\beta_i^1, \beta_i^2, \cdots, \beta_i^n\}$$

定义 2　B 细胞网络为 B 细胞构成的集合:

$$bnet = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m\}$$

定义 3　抗原集为待分析的数据集:

$$ag = data = \{(y_i, \boldsymbol{X}_i)\} = \{(y_i, [x_i^1, x_i^2, \cdots, x_i^n]^{\mathrm{T}})\}$$

其中:$data$ 是待分析数据集;$\boldsymbol{X}$ 为输入向量;y 为输出数据;$i \in [1,2,\cdots,s]$,s 为数据个数。

定义 4　抗原 i 与 B 细胞 j 的符合程度为:

$$af_{ij}^1 = \begin{cases} |y_i - \boldsymbol{b}_j * \boldsymbol{X}_i|, & |y_i - \boldsymbol{b}_j * \boldsymbol{X}_i| > \varepsilon \\ 0, & |y_i - \boldsymbol{b}_j * \boldsymbol{X}_i| \leqslant \varepsilon \end{cases} \tag{1}$$

其中,af^1 越小,两者符合程度越高,值为 0 表示完全符合,完全符合的数据称为该细胞的有效数据。ε 是判定是否完全符合的阈值。

定义 5　B 细胞 i 与 j 之间的近似程度为:

$$af_{ij}^2 = \begin{cases} 1, & |\boldsymbol{b}_i - \boldsymbol{b}_j| > \delta \\ 0, & |\boldsymbol{b}_i - \boldsymbol{b}_j| \leqslant \delta \end{cases} \tag{2}$$

其中:af^2 值为 0 表示两者近似,δ 是判断两个 B 细胞是否近似的阈值。

具体算法流程如下:

步骤 1　输入抗原集 ag,确定回归模型参数数目 n;确定初始网络 $bnet$ 的细胞个数 m,随机生成 $bnet$;确定阈值参数 ε、δ 及有效模型数据长度阈值 L;确定结束标准。

步骤 2　补充 p 个新生 B 细胞。具体方法是:随机选取 q 个抗原,利用常规回归算法得出一组回归参数,作为新细胞。补充到 $bnet$ 中。

步骤 3　以 $bnet$ 中每个 B 细胞为一个子类,将 ag 分成 m 个子类。

1) 按式(1) 分别计算 ag 中所有数据与 $bnet$ 中所有 B 细胞的符合程度,得出符合度矩阵 $\boldsymbol{AF}^1 = \{\boldsymbol{AF}_1^1, \boldsymbol{AF}_2^1, \cdots, \boldsymbol{AF}_s^1\}$。其中 $\boldsymbol{AF}_i^1 = [af_{i1}^1, af_{i2}^1, \cdots, af_{im}^1]^{\mathrm{T}}$,为第 i 个抗原与所有 B 细胞符合程度值构成的列向量。

2) 针对每一个抗原 i,根据其符合程度向量 $\boldsymbol{AF}_i^1$,划分该抗原归属 B 细胞种类。将 $\boldsymbol{AF}_i^1$ 按 af_{ij}^1 的值由低到高进行排序,如果 $af_{ij1}^1 = af_{ij2}^1$,则依据第 j 类拥有数据量的多少,从高到低排序。假设 af_{ij}^1 为排序后向量的首个值,则该抗原 i 归属 B 细胞 j 类。

3) 剔除拥有数据量小于阈值的 B 细胞类。根据步骤 3 中的 2) 的分类,计算每一类数据集的数据个数,如果小于阈值 L,则认为该数据集不构成有效数据类,予以剔除。

步骤 4　根据步骤 3 的分类结果,进行免疫网络 $bnet$ 的修改与更新。

1) 对由步骤 3 得到的各有效数据类分别进行常规鲁棒回归计算,得出每个数据类的回归模型,形成更新的 B 细胞集 $bnet'$。

2) 通过网络抑制排除 $bnet'$ 中的冗余细胞。利用式(2) 计算网络 $bnet'$ 中所有细胞间的相似程度。对于相似细胞组,细胞对应的数据类越大,则认为该细胞受正向激励越大,生存能力越强,因此保留对应数据类最大的 B 细胞,其他的作为冗余细胞被抑制掉。经过网络抑制后的细胞集作为下一代 $bnet$。

步骤 5　判断是否结束迭代。结束标准是看 $bnet$ 是否已经连续经历 t 次迭代而没有发生变化。不满足结束条件返回步骤 2;满足则结束迭代。转步骤 6。

步骤 6　提取 $bnet$ 的内核作为模型集输出。即对 $bnet$ 中各细胞所属数据进行一致性检验,排除有效数据数目小于阈值 L 的细胞,剩余部分为 $bnet$ 的内核。

3　仿真实验

本文选择图像处理中常见的直线参数提取问题作为算例来比较算法的性能优劣。即根据输入输出数据集提取如下模型集参数:

$$y = b_i^0 + b_i^1 x; \quad i \in [1,2,\cdots,N]$$

其中 N 为模型个数。

将本文免疫回归方法与目前典型的“聚类 + 回归”“逐次提取”两种方法进行比较。其中免疫回归中的单模型提取方法采用经典的 M 估计。为了全面比较,“聚类 + 回归”“逐次提取”两种方法中的单模型提取算法分别采用 M 估计、LMS、LTS 和 RANSAC 算法来进行效果比较。“聚类 + 回归”中的聚类方法采用经典的 K 聚类,聚类数假定已知为模型数。

测试数据随机产生。具体比较 3 种情况:

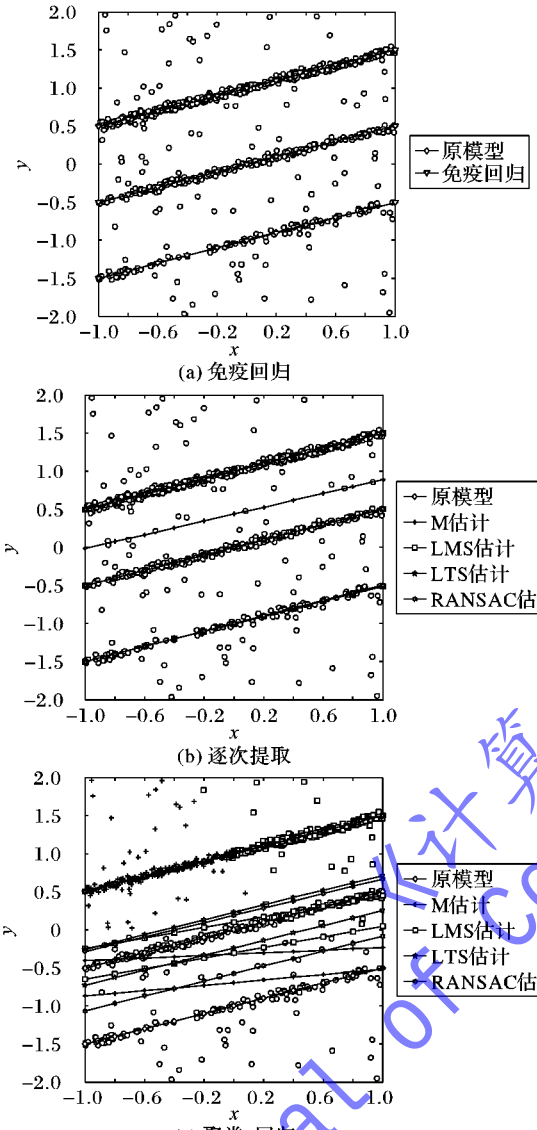
1) 情况 1。模型数目较少,按有效数据多少不同模型有主次之分。

2) 情况 2。模型数目较少,不同模型数据容易被准确分类。

3) 情况 3。模型数据较多,不同模型的数据交叉混叠在一起,不易被准确分类。

情况 1　选择模型 1,2,3 的参数分别是 $(b_i^0, b_i^1) = (-1, 0.5), (0, 0.5), (1, 0.5)$。在 $x \in [-1,1]$,按模型 1 数据数目为 400,模型 2 为 200,模型 3 为 100,随机产生测试数据。每个数据存在偏差,其分布符合高斯分布,标准差为 0.03。另外在数据集中加入随机生成的 100 个离群数据,取值范围为 $x \in [-1,1]$,$y \in [-2,2]$。合计 800 个数据。将离差在 0.03 以内的数据作为模型有效数据,拥有 70 个有效数据以上的模型被认为是有效模型。在逐次提取过程中,如果剩下 100 个数据,则认为提取过程结束。免疫回归结束标准为 $t = 10$;LMS、LTS 采样数为 10000;RANSAC 迭代次数为 1000;以欧氏距离作为模型误差测度,在 0.1 以内,则认为发现了该模型。为了测试算法稳定性,每种算法进行 100 次蒙特卡洛计算。取均值作

为比较结果。图 1(a)～(c)为 3 种方法所得的模型集,其中“o”表示测试数据。为便于分析,图 1(c)中测试数据根据聚类结果用不同的符号加以表示。100 次运算的结果比较如表 1 所示。



(a) 免疫回归

(b) 逐次提取

(c) 聚类+回归

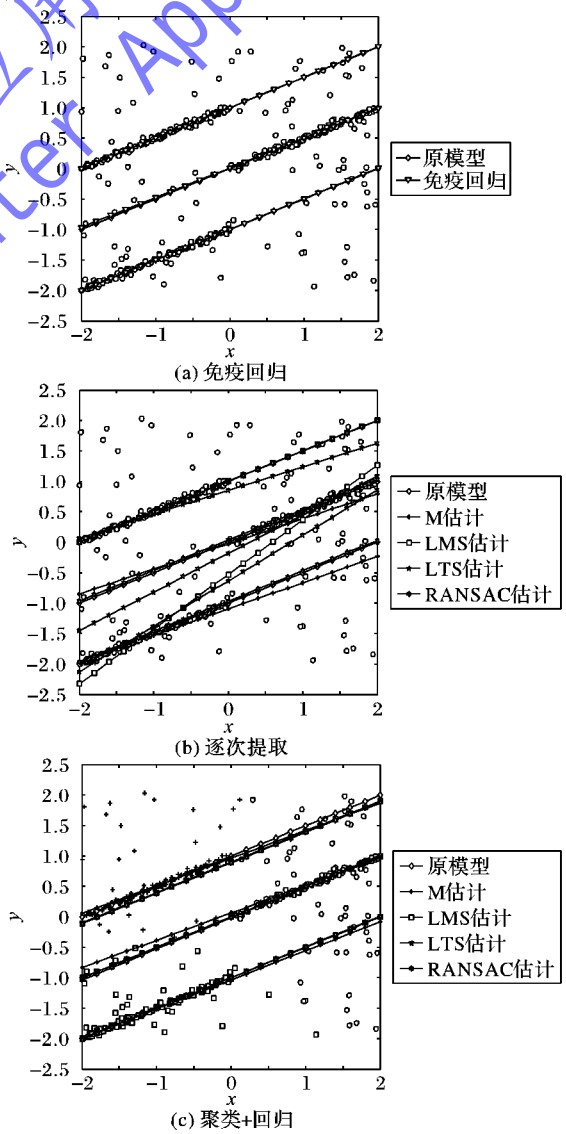
图 1 情况 1 的各方法性能比较

表 1 情况 1 的 3 种方法法性能比较

算法		计算时间/s	模型发现率/% 模型 1	模型 2	模型 3	模型误差均方根 模型 1	模型 2	模型 3
免疫回归算法		0.1762	100	100	100	0.0071	0.0070	0.0065
逐次提取算法	M 估计	0.0081	0	0	100	0.5537	0.4486	0.0070
	LMS	1.8112	100	100	100	0.0118	0.0128	0.0126
	LTS	1.9469	100	100	100	0.0036	0.0049	0.0071
	RANSAC	27.5026	99	100	97	0.0358	0.0293	0.1734
聚类+回归算法	M 估计	0.0048	100	11	22	0.0059	0.6349	0.5642
	LMS	1.3479	100	68	22	0.0121	0.4288	0.8628
	LTS	1.4797	100	77	23	0.0113	0.4746	0.8685
	RANSAC	10.9559	65	49	43	0.1229	0.5336	0.7497

情况 1 的测试结果分析 本例中,3 个模型的有效数据在数据量上有主次之分,按模型 1、模型 2、模型 3 的顺序逐次抽取,则每个模型在抽取过程中,其离群点的比例在 50% 左右,因此适合逐次抽取。不过,由于数据聚集特征不明显,不容易被准确分类,从图 1(c)来分类结果看,模型 1 的数据多归于同一类,模型 2 和模型 3 的数据相互交叠,没有有效分开,因此不适合采用“聚类 + 回归”的方法。从表 1 的模型发现率来看,“逐次提取”的效果确实远比“聚类 + 回归”要好。当然,在“逐次提取”诸方法中,M 估计效果不佳,原因是其崩溃点远小于其他 3 种,在含 50% 离群点的条件下,仍然难以得到较好的效果。“聚类 + 回归”诸方法中,模型 1 的发现率较高,原因是其数据被准确分类出来了。相比之下,本文算法虽然也采用 M 估计作为单模型提取方法,但是由于采用了反复修正的多次“分类 + 回归”的模式,既有效利用了 M 估计计算速度快的特点,也保证了回归分析的准确率。

情况 2 模型与参数设置与情况 1 相同,不同的是模型 1 ～3 的测试数据均为 100 个,不从属任何模型的离群数据 100 个,总计测试数据 400 个。模型 1 和模型 3 的 x 取值范围是 $x \in [-2,0]$,模型 2 的 x 取值范围是 $x \in [0,2]$。同样每种算法进行 100 次蒙特卡洛计算。取均值作为比较结果。图 2(a)～(c)为 3 种方法所得的模型集,表 2 为 100 次运算的结果。



(a) 免疫回归

(b) 逐次提取

(c) 聚类+回归

图 2 情况 2 的各方法性能比较

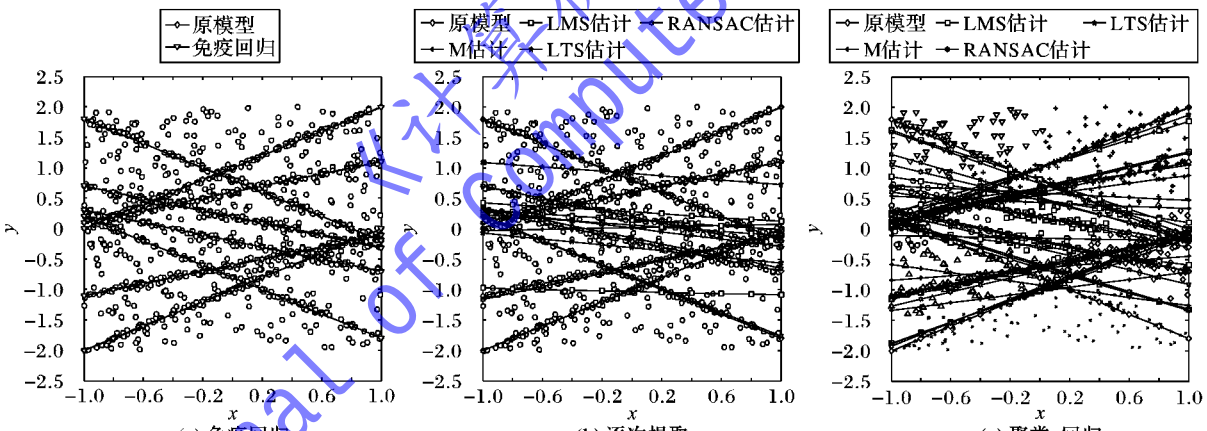
情况 2 的测试结果分析　从数据集的构成来看,本例中各模型没有主次之分,因此,在逐次剥离的过程中,先剥离的模型面对大量的离群点,其回归效果肯定不好。从表 2 的模型发现率来看,逐次提取法中,M 估计、LMS、LTS 效果都很差,只有 RANSAC 由于崩溃点高仍然维持较好效果。另外,从图 2(c)来看,由于 3 个模型数据能够被聚类方法较为准确地分开来,因此,本例适合采取“聚类 + 回归”的方法。从表 2 的模型发现率来看,“聚类 + 回归”的 4 种方法,除 RANSAC 因为估计精度不够,造成部分模型丢失外,效果都不错。从图 2(a)和表 2 来看,本文方法的模型发现率和回归精度都不错,计算时间虽然比 M 估计长,但是远优于其他方法。

表 2　情况 2 的 3 种方法性能比较

算法		计算时间/s	模型发现率/%			模型误差均方根		
			模型 1	模型 2	模型 3	模型 1	模型 2	模型 3
免疫回归算法		0.1132	100	100	100	0.0154	0.0115	0.0172
逐次提取算法	M 估计	0.0056	0	0	0	1.0482	0.1577	0.5431
	LMS	0.8037	0	0	100	0.6739	0.6279	0.0209
	LTS	0.8714	0	0	100	0.6181	0.6435	0.0099
	RANSAC	6.4809	98	94	94	0.1511	0.2304	0.2518
聚类 + 回归算法	M 估计	0.0033	100	100	100	0.0084	0.0090	0.0085
	LMS	0.9146	100	100	100	0.0203	0.0161	0.0160
	LTS	1.0242	100	100	100	0.0192	0.0159	0.0162
	RANSAC	4.9621	79	95	97	0.0780	0.0615	0.0559

情况 3　选择模型 1 ~ 8 的参数分别是$(b^0,b_i^1)=(1,1)$,$(-1,1)$,$(0.6,0.5)$,$(-0.6,0.5)$,$(0.2,-0.5)$,$(-0.2,-0.5)$,$(0.8,-1)$,$(-0.8,-1)$,在$x\in[-1,1]$,按每个模型 100 个测试数据随机产生测试数据,数据偏差分布为高斯分布,标准差 0.03。在$x\in[-1,1]$,$y\in[-2,2]$内随机生成的 400 个离群数据加入数据集,合计 1200 个数据。模型有效数据、有效模型定义与前两例相同,在逐次提取过程中,如果剩下 400 个数据,则认为提取过程结束。免疫回归结束标准仍为$t=10$。每种算法进行 100 次蒙特卡洛计算,取均值作为比较结果。图 3(a) ~ (c)为 3 种方法所得的模型集,表 3 为 100 次运算的结果比较。

情况 3 的测试结果分析　本例中,首先,由于模型较多,采样数据相互交叉混叠,不易准确分类。从图 3(c)的分类结果来看,8 个类别的数据没有按 8 个模型有效分开,因此,根据这个分类进行模型抽取效果肯定不好。实际上图 3(c)和表 3 的模型发现率来看,4 种方法抽取的模型与实际模型集都相差较远。其次,由于各模型在采样数据上也没有主次之分,离群点比例远高于 50%,因此崩溃率低的算法在逐次提取过程中难以发挥作用。从图 3(b)和表 3 的模型发现率来看,除 RANSAC 仍保持较高的识别率外,其他方法效果都很差。相比之下,本文方法仍然能够以较高的精度发现模型集。其模型发现率和估计精度也高于 RANSAC,尤其是计算时间远低于 RANSAC。



(a) 免疫回归　(b) 逐次提取　(c) 聚类+回归

图 3　情况 3 的各方法性能比较

表 3　情况 3 的 3 种方法性能比较

算法		计算时间/s	模型发现率/%								模型误差均方根							
			模型 1	模型 2	模型 3	模型 4	模型 5	模型 6	模型 7	模型 8	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6	模型 7	模型 8
免疫回归算法		0.5094	98	91	98	99	100	100	100	100	0.0911	0.1463	0.0743	0.0469	0.0078	0.0125	0.0081	0.0100
逐次提取算法	M 估计	0.0058	0	0	0	0	0	0	0	0	1.4509	1.4056	0.8184	0.7695	0.5337	0.5089	1.2861	1.2453
	LMS	1.6873	0	0	0	0	0	0	0	0	1.2188	1.4167	0.8849	0.8580	0.4262	0.3485	1.0763	1.0870
	LTS	1.7767	0	0	0	0	0	0	0	0	1.6649	1.2135	1.0251	0.9360	0.2624	0.2629	1.0428	0.9170
	RANSAC	53.3947	84	93	96	98	95	78	82	97	0.2581	0.1727	0.1316	0.0995	0.1485	0.2209	0.3314	0.1591
聚类 + 回归算法	M 估计	0.0103	0	3	0	1	0	2	0	0	1.0302	1.0238	0.5198	0.5193	0.4778	0.4930	0.9395	0.9493
	LMS	2.7066	4	40	15	50	2	6	12	49	0.4219	0.2972	0.2330	0.1470	0.7093	0.6022	0.4828	0.4311
	LTS	3.0234	7	78	39	78	9	8	82	82	0.3150	0.2610	0.2296	0.1066	0.7222	0.6140	0.4534	0.4150
	RANSAC	8.4563	23	69	44	63	35	25	55	58	0.9325	0.2264	0.5769	0.2880	0.4662	0.5017	0.6900	0.6322

从上述算例的结果来看:“逐次提取”和“聚类 + 回归”两种方式受数据集的结构影响较大。前者需要不同模型的数据量保持一定的比例,以保证每次模型提取都满足离群点比例低于算法崩溃点;后者则要求不同模型的数据能够被较为准

确地被分类。从结果看,由于崩溃点高,采用 RANSAC 对数据集进行逐次提取准确率较高。不过也可以看出,RANSAC 方法在几种算法中是最耗时的,而且随着数据量的增加,算法所需时间迅速增加,这也是限制该方法在大数据量、多模型条件下应用的一个重要原因。相比之下,无论数据集是否适合分类或满足特殊的比例关系,本文方法均能够较为准确地获得模型集各模型参数,计算时间短,具有更好的多模型数据集回归分析能力。

4 结语

针对多模型数据集的鲁棒回归分析问题,采取模型逐次提取的模式则对回归算法抗离群数据干扰的能力要求很高;采取先将数据分类再分别回归分析的模式,则对聚类算法准确分类数据集的能力要求很高。本文将免疫聚类方法与鲁棒回归结合起来,采用免疫网络记忆临时模型集,通过判断数据是否符合模型来实现数据集的分类,通过比较模型识别数据的能力来进行有效模型的筛选,通过“尝试性分类-回归分析-修正分类-再回归分析”的启发式迭代过程,不断逼近最优模型集。该方法由于不需要一次性完成回归分析或者聚类分析,因而更容易实现。仿真结果表明:该方法在数据集较为复杂的情况下,用较少的计算时间,仍然能够获得满意的模型集抽取效果。

参考文献:

[1] LI X H, SHEN H F, ZHANG L P. Dead pixel completion of aqua MODIS band 6 using a robust M-estimator multiregression [J]. Geoscience and Remote Sensing Letters, 2014, 11(4): 768 – 772.

[2] ROUSSEE P J. Least median of squares regression [J]. Journal of the American Statistical Association, 1984, 79(388): 871 – 880.

[3] JIE S, ZHOU S K, CHELLAPPA R. Robust height estimation of moving objects from uncalibrated videos [J]. IEEE Transactions on Image Processing, 2010, 19(8): 2221 – 2232.

[4] KIM J, KRISHNAPURAM R, DAVE R N. Application of the least trimmed squares technique to prototype-based clustering [J]. Pattern Recognition, 1996, 17(6): 633 – 641.

[5] SHEN F M, SHEN C H, van DEN HENGEL A, *et al*. An approximate least trimmed sum of squares fitting and applications in image analysis [J]. IEEE Transactions on Image Processing, 2013, 22(5): 1836 – 1847.

[6] CHUM O, MSATAS J. Optimal randomized RANSAC [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(8): 1472 – 1482.

[7] XU M, LU J. Distributed RANSAC for the robust estimation of three-dimensional reconstruction [J]. Computer Vision, 2012, 6(4): 324 – 333.

[8] CHIN T J, YU J, DAVID S. Accelerated hypothesis generation for multistructure data via preference analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 625 – 638.

[9] CHEN H F, PETER M, DAVID E, *et al*. Robust regression for data with multiple structures [C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2001: 1069 – 1075.

[10] SHAPOSHNYK V, VILLA A E P, AKSENOVA T. Advances in structural modeling robust to outliers in explanatory and response variables [C]// Proceedings of the 2010 International Joint Conference on Neural Networks. Piscataway: IEEE Press, 2010: 1 – 8.

[11] CHERKASSKY V, MA Y Q. Multiple model regression estimation [J]. IEEE Transactions on Neural Networks, 2005, 16(4): 785 – 797.

[12] NURUNNABI A A M, NASSER M. Regression diagnostics for multiple model step data [C]// Proceedings of the 2009 IEEE International Conference on Digital Image Processing. Piscataway: IEEE Press, 2009: 85 – 89.

[13] GRAAFF A J, ENGEIBRECHT A P. Clustering data in an uncertain environment using an artificial immune system [J]. Pattern Recognition Letters, 2011, 32(2): 342 – 351.

(上接第 2259 页)

[9] KAUSHIK R T, BHANDARKAR M. Green HDFS: towards an energy-conserving, storage-efficient, hybrid Hadoop compute cluster [C]// Proceedings of the 2010 International Conference on Power Aware Computing and Systems. Piscataway: IEEE Press, 2010: 1 – 9.

[10] KAUSHIK R T, BHANDARKAR M, NAHRSTEDT K. Evaluation and analysis of green HDFS: a self-adaptive, energy conserving variant of the Hadoop distributed file system [C]// Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science. Piscataway: IEEE Press, 2010: 274 – 287.

[11] DONG J, CHEN W, WU H, *et al*. Load balancing study in cloud storage based on dynamic replica technology [J]. Application Research of Computers, 2012, 29(9): 3422 – 3425. (董继光, 陈卫卫, 吴海佳, 等. 基于动态副本技术的云存储负载均衡研究[J]. 计算机应用研究 2012, 29(9): 3422 – 3425.)

[12] NITESH M, NANDURI R, VARMA V. Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework [J]. Future Generation Computer Systems, 2011, 28(1): 119 – 127.

[13] ZHU Q, CHEN Z, TAN L, *et al*. Hibernator: helping disk arrays sleep through the winter [C]// Proceedings of the 20th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2005: 177 – 190.

[14] LIAO B, YU J, SUN H, *et al*. Energy-efficient algorithms for distributed storage system based on data storage structure reconfiguration [J]. Journal of Computer Research and Development, 2013, 50(1): 3 – 18. (廖彬, 于炯, 孙华, 等. 基于存储结构重配置的分布式存储系统节能算法[J]. 计算机研究与发展, 2013, 50(1): 3 – 18.)

[15] LIAO B, YU J, ZHANG T, *et al*. Energy-efficient algorithms for distributed file system HDFS [J]. Chinese Journal of Computers, 2013, 36(5): 1047 – 1064. (廖彬, 于炯, 张陶, 等. 基于分布式文件系统 HDFS 的节能算法[J]. 计算机学报, 2013, 36(5): 1047 – 1064.)