文章编号: 1009 - 2552(2018)05 - 0133 - 06　　DOI: 10.13274/j.cnki.hdzj.2018.05.030

Logistic 回归算法研究与实现

滕　文

(陕西国际商贸学院信息与工程学院，西安 712000)

摘　要: 针对目前标记噪声普遍在数据集中出现的这一现象，研究了一种新的模型，称为鲁棒逻辑回归模型。该模型以传统的贝叶斯逻辑回归模型为基础，在分类器中加入标记转换概率来应对可能出现的标记噪声。同时在模型中运用了正则化的方法，使分类器在拟合数据与变量选择间保持平衡。实验中分别用到了合成数据集和真实的数据集，从而对鲁棒逻辑回归模型在分类问题中具有的预测能力和变量选择能力以及对标记噪声的鲁棒性进行验证，再与传统的模型进行比较。结果表明在面对含有标记噪声的数据时，由鲁棒逻辑回归模型训练产生的分类器有更低的误分类率，在变量选择方面也更准确。

关键词: Logistic 回归; 标记噪声; 稀疏性; 鲁棒性

中图分类号: TP302.7　　**文献标识码:** A

Research and implementation of Logistic regression classifier algorithmic

TENG Wen

(School of Computer Science, Shaanxi Institute of International Trade and Commerce, Xi'an 712000, China)

Abstract: Labelling errors are common in the microarray data sets. A new model is studied on. It is called robust logistic regression model. In the new model, which is constructed on the basis of the Bayesian logistic regression model, added label-filpping probability in the classifier to cope with the labelling errors. It regularizes in the objective function, balances the classifier between the over-fitting problem and the variable selection capability . The experiments use synthetic data sets and real data sets, testing predictive ability, variable selection ability and robustness against labelling errors of robust logistic regression model, compared with the Bayesian logistic regression model. It turns out that when labelling errors are in data sets, the classifier trained by the robust logistic regression model has lower misclassication error and a more accurate estimation of the parameter.

Key words: logistic regression; labelling errors; sparsity; robustness

0　引言

随着社会的高速发展，生活节奏的加快，长期处于压力之下的人们就会养成许多不良的习惯，这可能会让人们得上慢性疾病。目前随着我国老龄化、城镇化、工业化的加快，慢性病患病和死亡率呈现加快增长的趋势。慢性病已经成为影响我国居民生活水平提高，阻碍经济社会发展的重大公共卫生问题和社会问题。因此有必要加强对慢性病的预防与研究工作，这就要用到统计分析方法。

本文采用的回归分析法是在大批的观察数据的基础上，使用数理统计方法成立因变量与自变量之间相关的回归关系函数表达式，其中的因变量是分类变量。通过 Logistic 回归模型来解决问题。

1　模型与算法

1.1　带标记噪声的鲁棒逻辑回归模型

要对含有噪声的训练集进行分类器的训练，需用到“鲁棒”这个词，以此将本文用到的模型与传统

收稿日期: 2017 - 08 - 28

作者简介: 滕文(1982 -)，男，硕士研究生，研究方向为计算机网络、大数据。

的逻辑回归模型区分开。鲁棒在这里是指分类器在面对噪声时尽可能不受其影响，维持它的基本特性，能做出正确的判断。

1.2 模型中的参数估计

经典的逻辑回归分类器是有明显的缺点，一个是过拟合问题，另一个是变量选择问题。过拟合是指逻辑回归分类器在对训练集的拟合往往能表现出较高的精度，而在训练集以外的数据集上的精度较差。如今过拟合问题不仅在逻辑回归分类器中存在，其他的数据分析模型也会有这种问题。变量选择问题是指分类器经过训练后得到的模型参数，这些参数大部分不为0。这些非零参数和设计的模型都有关系，也可以说成不具有稀疏性。而实际上，与模型有关的参数的数目不会太多，这样训练后大部分参数为0，这样不仅有利于问题的解释，还能简化计算。为了解决这两个问题，可以利用正则化这一方法。

在流行病研究中，基因数据往往是高维的，含有很多特征，多达上千个。但是与模型有关的特征数目却很少，往往只有几十个。这里可以用正则化来解决，让模型中的参数变的稀疏。因此对前面的目标似然函数可以加入 $L1$ 正则化公式，得到如下新的目标函数:

$$\max_{w}\sum_{n=1}^{N}\log p(\tilde{y}_n \mid x_n, w) - \sum_{i=1}^{m}\alpha_i \mid w_i \mid \qquad (1)$$

其中，m 是特征的数目，α_i 是正则化参数，α_i 的作用是让模型在具有良好的拟合性与有较少的参数数目间平衡。公式(1)中的目标函数不可导，为了对其求导，可以采用一个很简单，让目标函数曲线变的平滑的方法。就是定义 $\mid w_i \mid \approx (w_i^2 + \gamma)^{1/2}$，同时令 $\gamma = 10^{-8}$。接下来就要求其中的正则化参数 α_i，一般可以采用交叉验证法。这样就会用到验证集，但是不能保证验证集的正确性，而且 α_i 的数量较多，这样导致误差较大。所以可以改写 $L1$ 正则化，目标函数改写成:

$$\max_{w}\sum_{n=1}^{N}\log(\tilde{y}_n \mid x_n, w) - \lambda \parallel w \parallel \qquad (2)$$

其中，λ 是正则化参数，其作用与上面用到的 α_i 相同。正则项定义为:

$$\parallel w \parallel = \sum_{d=1}^{m} w_d \qquad (3)$$

接下来对正则化参数 λ 进行定义，当然这里不能用交叉验证法，不仅耗时较多，而且不能保证验证集是否标记正确。于是采用贝叶斯正则化方法，这种方法不包含交叉验证，而且对正则化参数的计算比较简单。

$$\lambda = \frac{n}{\sum_{i=0}^{n} \mid w_i \mid} \qquad (4)$$

式中，n 一定不会大于数据集的维数 m，因为 n 代表非零参数的个数，也就是 $w_i \neq 0$。

接下来就要估计模型的参数 w 和标记转换概率矩阵。以前有人用到过一个简单但是很有效的算法，这个算法使用 Gauss-Seidel 迭代法以及坐标下降法，来优化稀疏逻辑回归模型产生的非平滑凸函数。但这里得鲁棒逻辑回归模型产生的目标函数是非凸的，所以要对这个算法进行一些修改。

首先对目标函数求导，得到 $F_i = \frac{\partial L(w)}{\partial w_i}$，需要注意这里参数 w 是 $m+1$ 维的，比训练集多一维，w_0 就是多出来的这一项，这是基本项，不需要对其正则化。

如果对参数进行优化，那么一次只能对一个参数优化，如何找到这个参数是个问题，这里就要用到最大违反值这个概念。最大违反值在这里指对目标函数影响最大的参数 w_i，首先对其优化。相应优化参数的违反值是这样定义的: ①如果 $i = 0$，其违反值为 $\mid F_i \mid$; ②如果 $w_i > 0\ (i > 0)$，违反值为 $\mid \lambda - F_i \mid$; ③如果 $w_i < 0(i > 0)$，违反值为 $\mid \lambda + F_i \mid$; ④如果 $w_i = 0(i > 0)$，违反值为 $\max(F_i - \lambda, -\lambda - F_i, 0)$。在找到与最大违反值对应的参数 w_i 后，就要对其优化。如果目标函数是凸函数，就能用梯度的方法来优化参数使目标函数取极值。但本文研究模型的目标函数是非凸的，梯度方法在这里是无效的。于是可以用搜索的方法，分别在 $(-\lim, 0)$ 和 $(0, \lim)$ 两个范围内搜索，找到能使目标函数取最大值得参数 w_i。

1.3 高维小样本对模型产生的影响

本文用到模型中的伽马表的估计是从给定的训练集中估计而得到的，这里就希望有足够多的训练样本提供给模型，这样能使鲁棒逻辑回归的优势发挥出来。但实际上慢性病研究中提供的与某个疾病有关的样本数量很少，难以满足模型的要求。这种情况下就要根据样本可能的误标记比例来预设伽马表，对误标记比例的估计也要根据以往实验中样本的表现来给出。只有这样，才能对模型中的目标函数进行优化。在优化过程中既可以保持伽马表不变，也可以让伽马表随着优化的过程而更新。由于本文中模型用到的数据集样本数目较少，只有几十个，而且维度较高，多达千维。两者相差过大，因此

该采用伽马表固定的鲁棒逻辑回归模型用于数据集的验证。

1.4 算法流程

本文中鲁棒逻辑回归模型的算法主要分成两个部分，第一部分是先找到需要优化的参数 w_i，第二部分是对参数 w_i 优化的过程。

算法 1: 主循环

输入: 训练数据集。把模型参数 W 都设为 0，目标函数中的正则化参数 λ 设为 0，

设置两个参数集合 G_{nz}，G_z，$G_{nz}=\{w_0\}$，$G_z=\{w_1,w_2,\cdots,w_m\}$，预设伽马表。

输出: 优化后的模型参数 W 和伽马表。

Step 1: 在集合 G_z 中寻找参数 w_i，其违反值最大。

Step 2: 在算法 2 中对参数 w_i 优化，再把其从集合 G_z 中删除，加入集合 G_{nz} 中。

Step 3: 如果集合 G_{nz} 中存在参数 w_j 其违反值最大，则对其优化，重复 Step2，否则转到 Step4。

Step 4: 更新正则化参数 λ 和标记转换概率。

Step 5: 如果集合 G_z 中的参数无违反值，算法结束。否则转到 Step1。

算法 2: 对参数优化

输入: 待优化参数 w_i，集合 G_z，G_{nz}。

输出: 优化完的参数 w_i。

Step 1: 假设 w_i 为 0，如果 w_i 满足优化条件则把参数从集合 G_{nz} 中删除再加入 G_z 中，算法结束，否则转到 Step2。

Step 2: 分别在区间 $(-1000,0)$ 和 $(0,1000)$ 搜索参数 w_i，使的目标函数最大化。

2 实验与结果

在本实验中，将本文用到的鲁棒逻辑回归模型，伽马表固定的鲁棒逻辑回归模型与传统的贝叶斯逻辑回归模型相比较，以此对将本文研究的模型进行验证。然而在验证之前，先用到对称的标记转换概率和非对称的标记转换概率这两个概念。对称的标记转换概率指所有的样本都有相同的概率被误标记，而非对称的标记转换概率指属于不同类的样本有不同的误标记概率。通常模型中含有非对称的标记转换概率会降低算法的实现效果，因为类的判定边界很可能被改变了，实际上非对称的标记转换概率是很常见的，于是实验中会用非对称的标记转换概率。

实验中将用到两个数据集，一个是合成的数据集，另一个是真实的 colon 数据集。在合成数据的实验中，样本是服从标准高斯分布的多维数据。首先预设一个参数向量 w，其中 $w_1=w_2=w_3=10/3$，$w_i=0, i>3$，然后依据判定函数将样本进行分类，分别为类别 1 和 2。这里创建两个不同的训练集，一个有 500 个样本，另一个有 100 个样本，分别称之为 train－500 和 train－100，对其训练时设置非对称标记转换概率为 30%。而每次用来测试的数据集大小固定为 100。这里样本的维数不是固定不变的，在 100 维到 1000 维间取随机值，合成数据上进行的实验将在不同的维度上进行。在真实数据集的实验中，样本是包含标记噪声的，有固定的标记转换概率，还有当用到这些数据时，需将它们标准化。

2.1 利用合成数据实验

在 train－500 中，每个类各有样本 250 个，同时在 train－100 中，每个类各有样本 50 个。接下来依据 30% 的非对称标记转换概率，可以对样本的标记进行这样的改变。在 train－500 中，对属于类别 1 的样本不进行改变，对属于类别 2 的样本中随机选取 75，将其标记改为 1。在 train－100 中，同样对属于类别 1 的样本不做改变，在属于类别 2 的样本中随机选取 15 个，将其标记改为 1。那么样本中就包含噪声了，实验依次在维数为 100，200，400，600，800，1000 的合成数据上进行，在每个维数上分别进行 100 次实验。

在 train－500 和 train－100 上得到的分类器的误分类率分别在表 1－2 中显示。从表 1 可以看出，在相同维度的情况下，鲁棒逻辑回归模型比贝叶斯逻辑回归模型有更低的误分类率，这足以体现了当训练集包含噪声时鲁棒逻辑回归模型的优势。而在表 2 中，这两个模型的误分类率没有差别。这是因为用来训练的样本数目本来就很少，只有 100 个，远不及表 1 中的 500 个。而用鲁棒逻辑回归模型进行分类器训练的过程中，需要对标记转换概率进行准确的估计，这就需要大量的训练样本使分类器变的更完美。但是把模型中的标记转换概率固定下来，称之为伽马表固定的鲁棒逻辑回归模型，经过这个模型训练后产生的分类器的误分类率就会小许多。这就可以认为当训练集样本数目较少时，应该把模

表 1 模型在 train－500 上进行训练后得到分类器的误分类率

维数	贝叶斯逻辑回归	鲁棒逻辑回归
100	0.18	0.03
200	0.19	0.02
400	0.26	0.00
600	0.21	0.02
800	0.18	0.07
1000	0.19	0.01

型中的标记转换概率固定下来，在算法优化过程中不对其进行更新。

表 2　模型在 train – 100 上进行训练后得到分类器的误分类率

维数	贝叶斯逻辑回归	鲁棒逻辑回归	伽马表固定的鲁棒逻辑回归
100	0.26	0.24	0.18
200	0.28	0.26	0.21
400	0.29	0.29	0.25
600	0.32	0.30	0.27
800	0.32	0.31	0.27
1000	0.33	0.31	0.28

在对模型的分类性能进行比较后，接下来就要评价其在变量选择方面的能力。这里仍然用鲁棒逻辑回归模型与贝叶斯逻辑回归模型两种模型，所用的训练集是 train – 500，并且含有 30% 的非对称标记转换概率。

由于起初对合成数据的标记所用到的是这样的一个参数向量 w，其中 $w_1 = w_2 = w_3 = 10/3$，$w_i = 0$，$i > 3$。那么模型经数据集训练后在变量选择上应该与最初的参数 w 相似，只选择第 1，2，3 个变量。从图 2 可以看出鲁棒逻辑回归模型在变量选择方面比较精确，而在图 1 看到贝叶斯逻辑回归模型受噪声的影响比较大，选择了太多错误的非零参数。以此认为鲁棒逻辑回归模型在变量选择方面优于贝叶斯逻辑回归模型。

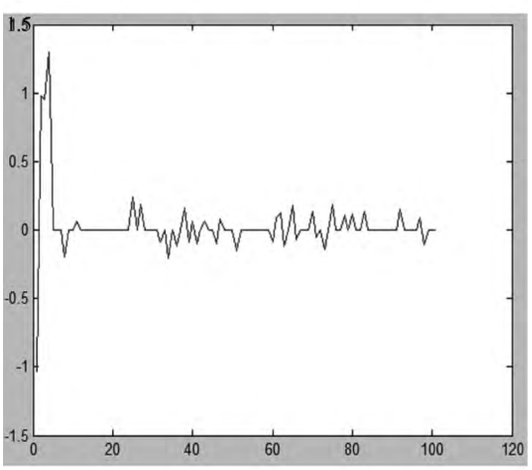

图 1　在贝叶斯逻辑回归模型上进行变量选择

2.2　利用真实数据实验

由于所用到的真实的数据集含有样本较少，这里应该采用交叉验证法。交叉验证法就本文来说就是在含 m 个样本的数据集上重复 m 次试验，每次选取一个不同的样本作为测试集，其余的 $m - 1$ 个样本作为训练集。这样能充分利用数据，来对模型进行验证。

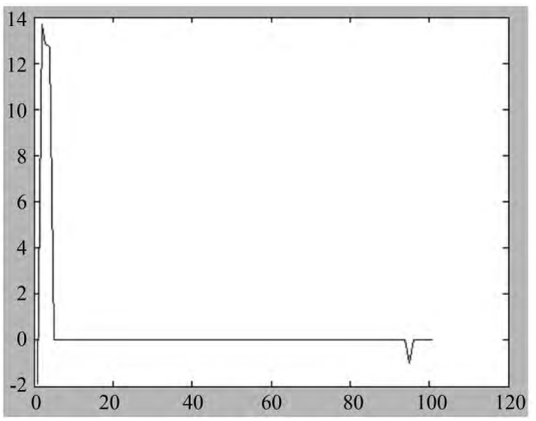

图 2　在鲁棒逻辑回归模型上进行变量选择

本实验用到了比较经典的 colon（结肠）数据集，其包含三个矩阵，分别是 x，y 和 fd。矩阵 x 是 62 × 2000 矩阵，包含 62 个样本，每个样本含有 2000 个不同的基因，基因是用数子来表示的。矩阵 y 是62 × 1 的类别矩阵，只包含 1，2 这两个数字。1 表示与 y 矩阵行对应的 x 矩阵样本没有患结肠癌，2 表示患有结肠癌。矩阵 fd 也是 62 × 1 矩阵，它只有 0，1 两个数，0 表示与 x 矩阵行对应的样本标记无误，1 则表示标记错误。当用交叉验证法进行测试时，一个样本作为测试集，其余 61 个样本作为训练集，这样重复 62 次，每个样本都有一次机会作为测试集。这样可以确保实验的公平性。

表 3 中给出了两种模型在相同的数据集上测试的结果。可以看到无论在 CRT 测试集还是在 CLN 测试集上，本文用到的模型比传统的模型有更低的误分类率，而且选择与结肠癌有关的基因数目更少。

表 3　用交叉验证法计算误分类率

模型算法	在 CRT 上的误分类率	在 CLN 上的误分类率	选择的基因数目（ ± 标准差）
贝叶斯逻辑回归模型	0.081	0.075	10.903 ± 0.393
伽马表固定的鲁棒逻辑回归模型	0.048	0，019	8.27 ± 0.45

注：对基因数目的选择是在 CRT 上进行的。

已经得出了本文用到的模型比传统的模型有更低的误分类率这一结论，接下来就是研究影响是否患有结肠癌的是那些基因。在两个不同的模型上进行实验，相互比较，观察本文用到的模型在判断方面是否真的可靠。

由于样本的维度非常大，有 2000 维，因此要从中选取与结肠癌有关的基因难度比较大。结合上面用到稀疏逻辑回归的算法，针对这 62 个样本，每次

随机从中选取 50 个样本作为训练集，得到的参数向量作为 w_1，重复执行 1000 次，这样会得到 1000 个不同的参数向量。最后把这些向量相加后取均值，则绝对值最大的维度其相应的基因即为某人患结肠癌影响最大的基因，基因的影响力随着其相应的参数绝对值的大小而变化着。

图 3 -4 分别显示的是本文用到的模型和传统的模型所呈现的结果。其中横坐标表示的是与基因号相对应的参数号（从 1 到 2000），纵坐标表示参数值。为了显示的更直观，这里并没有对参数取平均值，而是直接将 1000 个参数向量相加。因为这两个图在外观上比较相似，因此大致可以认为本文的模型在优化参数方面的能力与传统的模型差不多。表 4 -5 具体给出了分别在两个模型上求出的影响结肠癌比较大的 10 个基因的序号，基因名称以及对应参数的平均值。两个表中出现的基因序号很相似，而且相同序号的参数均值差别也不大。

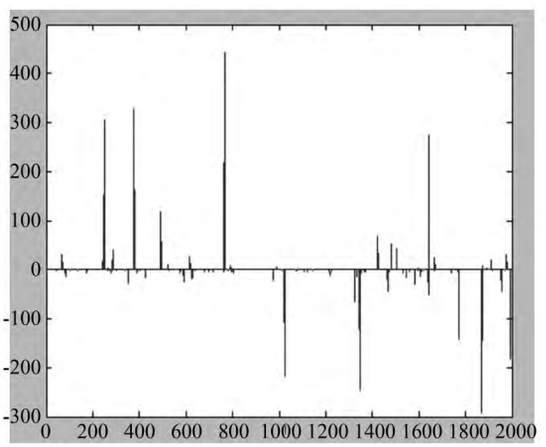

图 3　伽马表固定的鲁棒逻辑回归模型上进行变量选择

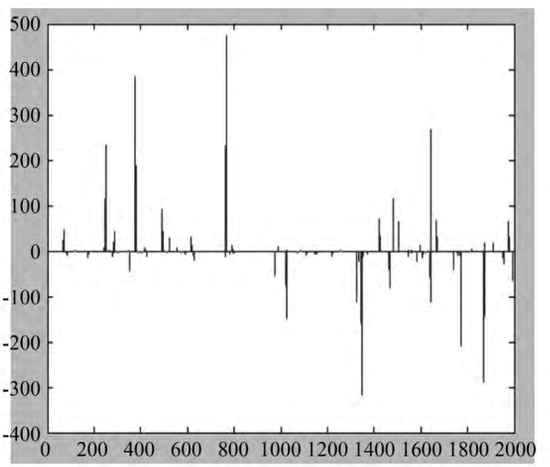

图 4　在贝叶斯逻辑回归模型上进行变量选择

表 4　通过贝叶斯逻辑回归算法的到相关性最大的 10 个基因

基因序号	基因名称	对应参数均值
765	Human cycteine-rich protein(CRP) gene, exons 5 and 6	0.476
377	H. sapiens mRNA for GCAP/ uroguanylin precursor	0.382
1346	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)	-0.317
1870	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDR IAL PRECURSOR	-0.289
1644	C4-DICABOXYLATE TRANSPORT SENSOR PROTENIN DCTB	0.266
249	Human desmin gene,complete cds	0.234
1772	COLLAGEN ALPHA 2 CHAIN	-0.209
1024	ATP SYNTHASE A CHAIN	-0.149
1482	Human spermidine synthase gene,complete cds	0.115
1641	Human enkephalin B gene,exon 4 and 3' flank and complete cds.	-0.112

表 5　通过固定伽马表的鲁棒逻辑回归模型得到相关性最大的 10 个基因

基因序号	基因名称	对应参数均值
765	Human cycteine-rich protein(CRP) gene, exons 5 and 6	0.443
377	H. sapiens mRNA for GCAP/uroguanylin precursor	0.328
249	Human desmin gene,complete cds	0.304
1870	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDR IAL PRECURSOR	-0.291
1346	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)	-0.245
1024	ATP SYNTHASE A CHAIN	-0.216
1993	Human hormone-sensitive lipase(LIPE) gene Completecds	-0.181
1772	COLLAGEN ALPHA 2 CHAIN	-0.141
493	MYOSIN HEAVY ALPHA 2(XI) CHAIN (Homo sapiens)	0.117

本实验通过合成数据和真实数据分别对鲁棒逻辑回归模型与贝叶斯逻辑回归模型进行误分类率与变量选择的比较，这些数据都包含标记噪声。从实验数据中可以看到在合成数据上，当样本数目较大时鲁棒逻辑回归分类器的误分类率更小，同时在变量选择上与传统的分类器比较相似。当样本数目较少时鲁棒逻辑回归分类器与传统分类器在判别误分类率上差别不大，如果把鲁棒逻辑回归模型中的伽马表固定下来，训练后的分类器就能有效的降低误

分类率。在真实数据的实验中由于样本数目较小，故把模型修改为伽马表固定的鲁棒逻辑回归模型，在与传统模型的比较中可以看出伽马表固定的鲁棒逻辑回归模型经训练后所得误分类率更小，在变量选择方面与传统的模型比较相似。这表明当数据包含标记噪声时，经鲁棒逻辑回归模型训练产生的分类器比贝叶斯逻辑回归模型的效果更好。

3 结束语

本文研究了一种以传统的贝叶斯逻辑回归模型为基础的分类器模型，称为鲁棒逻辑回归模型。经过上面的实验表明当训练集中含有标记噪声时，通过鲁棒逻辑回归模型产生的分类器性能优于传统的分类器，确切的说就是误分类率更低，尤其当训练集中的标记转换概率是非对称时更明显。同时由鲁棒逻辑回归模型学习产生的分类器在变量选择方面也十分有效。值得注意的是，鲁棒逻辑回归模型需要估计标记转换概率以及与模型参数，这就需要大量的训练集来使鲁棒逻辑回归模型的优势发挥出来。在本文的实验中当训练集数目比较少时，用鲁棒逻辑回归模型产生的分类器并不能明显降低误分类率，但是把模型中的伽马表固定后，误分类率就能大幅的降低，这表明在面对少量数据时需将鲁棒逻辑回归模型中的标记转换概率固定下来。

参 考 文 献:

[1] 李宏东，姚天翔. 模式分类 [M]. 北京: 机械工业出版社，2013.

[2] Bootkrajang J，Kaban A. Classification of mislabelled microarrays using robust sparse logistic regression [J]. Bioinformatics，2013，29(7)：870 - 877.

[3] Malossini A，Blanzieri E，Ng R T. Detecting Potential Labeling Errors in Microarrays by Data Perturbation [J]. Bioinformatics，2015，22(17)：2114 - 2121.

[4] Cawley G C，Talbot N L C. Gene Selection in Cancer Classification using Sparse Logistic Regression with Bayesian Regularization [J]. Bioinformatics，2014，22(19)：2348 - 2355.

[5] Bielza，C，Robles V，Larranagaa P. Regularized logistic regression without a penalty term: An application to cancer classification with microarray data [J]. Expert Systems with Applications，2014，38(5)：5110 - 5118.

[6] 李锐，李鹏，曲亚东，等. 机器学习实战 [M]. 北京: 人民邮电出版社，2015.

[7] 范明，昝红英，牛常勇. 机器学习导论 [M]. 北京: 机械工业出版社，2014.

[8] 申富饶，徐烨，郑俊，等. 神经网络与机器学习 [M]. 北京: 机械工业出版社，2015.

[9] 邓乃扬，田英杰. 支持向量机—理论、算法与拓展 [M]. 北京: 科学出版社，2014.

[10] 范明，柴玉梅，昝红英. 统计学习基础 [M]. 北京: 电子工业出版社，2014.

[11] Ayalew L，Yamagishi H. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains，Central Japan [J]. Geomorphology，2015，65(1/2)：15 - 31.

[12] King G，Zeng L. Logistic regression in rare events data [J]. Political analysis，2014，9(2)：137 - 163.

[13] Keating K A，Cherry S. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management，2014，68(4)：774 - 789.

[14] Zhu Ji，Hastie T. Classification of gene microarrarys by penalizad logistic regression [J]. Biostatistics，2014，5(3)：427 - 443.

[15] Genkin A，Lewis D，Madigan D. Large-scale Bayesian logistic reggression for text categorization [J]. Technometrics，2015，49(3)：291 - 304.

[16] Friedman J，Hastie T，Tibshirani R. Regularization paths for generalized linear models via coordinate descent [J]. Journal of statistical software，2015，33(1)：1 - 22.

责任编辑: 薛慧心

(上接第 132 页)

[2] Myron Flickner，Harpreet Sawhney，Wayne Niblack. Query by Image and Video Content [M]. Elsevier Inc.：2002.

[3] Brodersen Kay H，Schofield Thomas M，Leff Alexander P，et al. Generative embedding for model-based classification of fMRI data. [J]. PL o S Computational Biology (Online)，2011，7(6)：e10002079.

[4] Marrocco Claudio，Molinara Mario，D'Elia Ciro，et al. A computer-aided detection system for clustered microcalcifications. [J]. Artificial Intelligence in Medicine，2010，50(1)：23 - 32.

[5] 柯丽，张雪，康雁. 乳腺肿块多级分形特征提取方法研究 [J]. 生物医学工程学杂志，2013(2)：278 - 282.

[6] 陈娜，蒋芸，邹丽，等. 基于判别式受限玻尔兹曼机的医学图像分类法 [J]. 计算机科学，2015(5)：315 - 319.

[7] 周晓明，马秋禾，肖蓉，等. 一种改进的 Canny 算子边缘检测算法 [J]. 测绘工程，2008(1)：28 - 31.

[8] 段瑞玲，李庆祥，李玉和. 图像边缘检测方法研究综述 [J]. 光学技术，2005(3)：415 - 419.

[9] 赵宏中，张彦超. 基于 Canny 边缘检测算子的图像检索算法 [J]. 电子设计工程，2010(2)：75 - 77，80.

[10] Puolamäki Kai，Fortelius Mikael，Mannila Heikki. Seriation in paleontological data using markov chain Monte Carlo methods [J]. PL o S Computational Biology (Online)，2006，2(2)：e6.

[11] Hinton Geoffrey E，Osindero Simon，Teh Yee-Whye. A fast learning algorithm for deep belief nets. [J]. Neural Computation，2006，18(7)：1527 - 1554.

[12] Dambre Joni，Verstraeten David，Schrauwen Benjamin，et al. Information processing capacity of dynamical systems. [J]. Scientific Reports，2012，2(4)：514.

[13] 赵永威，李婷，蔺博宇. 基于深度学习编码模型的图像分类方法 [J]. 工程科学与技术，2017(1)：213 - 220.

[14] Yong-ping DU，Chang-qing YAO，Shu-hua HUO，et al. A new item-based deep network structure using a restricted Boltzmann machine for collaborative filtering [J]. Frontiers of Information Technology& Electronic Engineering，2017(5)：658 - 666.

[15] Hinton Geoffrey E. Training products of experts by minimizing contrastive divergence. [J]. Neural Computation，2002，14(8)：1771.

责任编辑: 肖滨