基于 RGMM 的离散基因表达数据关联规则挖掘

黄 睿
(湖南信息职业技术学院计算机系 湖南 长沙 410200)

摘 要 由于具有良好的可解释性,关联规则在基于疾病诊断的基因表达数据中表现出优越性,然而,高维基因表达数据中的大量规则阻碍了它的应用。为了缓解这个问题,提出正则化高斯混合模型 RGMM(Regularized Gaussian Mixture Model),根据最小描述长度框架,挖掘离散化模型复杂度及信息丢失准则,通过离散化连续的基因表达数据,缓解监督方法中的过拟合现象,并且改善无监督方法中的一些缺点。在六个分类数据集上的大量实验验证了所提方法的有效性。实验结果表明,与其他几种最先进的方法相比,所提的 RGMM 方法在现实的基因表达数据集中更具实用性。

关键词 离散化 基因表达数据 正则化高斯混合模型 关联规则 数据挖掘

中图分类号 TP399 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2014.09.047

DISCRIMINATION GENE EXPRESSION DATA ASSOCIATION MINING BASED ON RGMM

Huang Rui
(*Department of Computer Science, Hunan College of Information, Changsha 410200, Hunan, China*)

Abstract Association rule shows its advantage in disease diagnosis-based gene expression data for its good interpretability. However, the large number of rules in high dimensional gene expression data blocks its application. To mitigate this problem, we present the regularised Gaussian mixture model (RGMM), it mines the complexity of discretisation model and information loss criterion according to minimal description length framework, by discretisedand continuous gene expression data it relieves the over-fitting phenomenon in supervision method and improves some shortcomings in unsupervised method. The effectiveness of the proposed method has been verified by the extensive experiments on six classification data sets. Experimental results show that the proposed RGMM has better practical applicability in real-life gene expression data sets compared with several latest approaches.

Keywords Discretisation Gene expression data Regularised Gaussian mixture model Association rule Data mining

0 引 言

由于关联规则的解释具有简单性,它在疾病诊断基因表达数据库的应用中非常有用[1]。然而,对于离散基因表达数据来说,这些数据集的关联规则却很容易受到组合展开问题的影响,很难从大量的候选者中识别出真实的规则[2]。因此,离散基因表达数据关联规则的挖掘成了当今的研究热点[3]。

现有的离散化方法可以分为监督离散化和无监督离散化。ChiMerge[4]、Khiops[5]、SDM[5]、FUSINTER[6] 及 CAIM[7]是典型的监督离散化方法。此外,在特殊应用中,还有几种其它的离散化方法,例如,文献[8]提出的应用在简短网络建设中的克拉克离散化方法,文献[9]提出的应用在贝叶斯网络分类器中的傅里叶离散化方法,文献[10]还提出了在分类规则框架下挖掘标签信息的分类规则法。等间距宽度方法[11]、等频率方法[12]、多元离散化方法[13]以及最近文献[14]提出的基于离散化方法的核密度估计法 KDE(Kernel Density Estimation),都是主流的无监督离散化方法。然而,无监督离散化方法将数据离散化为很多区间,大大地增加了离散化数据分析的难度[15,16]。

基于上述分析,为了缓解监督方法中过拟合问题以及改善无监督方法中的一些缺点,提出了一种离散的正则化高斯混合模型 RGMM,假定样本符合混合的高斯分布,而且每个样本最大概率地在一个高斯分布中执行,实验结果验证了所提方法的有效性及可靠性。

1 方法提出

1.1 正则化高斯混合模型

给出了一个基因表达标准 x 及与它相应的离散间距 $t_1, t_2, \cdots, t_k$,离散化过程的运算如下:如果 $x > t_{j-1}$ 且 $x < t_j$,那么 x 离散化为第 j 个状态。

高斯混合模型可由一组高斯分布 G 来表示,第 j 个分量 G_j 是一个高斯分布 $G(\mu_j, \delta_j)$,其中 μ_j 是平均值,δ_j 是方差,所以第 j 个分量的不确定度是高斯分布的熵。其定义如下:

收稿日期:2013-05-25。**黄睿**,讲师,主研领域:数据挖掘,生物信息学。

$$EntopyG = 1n(2\pi e\delta_j^2)/2 \quad (1)$$

因为假设高斯混合模型中每个分量都相互独立的,所以高斯混合模型 G 的熵则为所有 m 个分量不确定度的总和。

$$Entopy(G) = \sum_{j=1}^{m} \ln(2\pi e\delta_j^2)/2 \quad (2)$$

假设每个样本都相互独立,那么全部的信息丢失就是所有 n 个样本的丢失,如下:

$$IL(S,G) = \sum_{i=1}^{n} 0.5\log\left(\frac{\delta_{c(s_i)}^2}{\delta_i^2}\right) = \sum_{i=1}^{n} KL(s_i,c(s_i)) + \frac{\mu_i^2 + \mu_{c(s_i)}^2 + \delta_i^2 - 2\mu_i\mu_{c(s_i)}}{2\delta_{c(s_i)}^2} \quad (3)$$

其中,$c(s_i)$是高斯混合模型样本 s_i的分量指数。

基于最小描述长度框架,最优离散化高斯混合模型可以通过优化式(4)中的目标函数得到,它是离散化复杂度以及离散化过程信息丢失的总和。从统计学习理论角度看,离散化模型中的熵是一种结构风险,而信息丢失是一种经验风险。

$$G_{opt} = \arg\min_{G}\{Entopy(G) + IL(S,G)\} \quad (4)$$

所提方法的整个算法过程如下算法 1 所示。

算法 1 离散化高斯混合模型算法过程

输入: S,原始数据;M,间距的最大量

结果: D,离散数据

```
δ = KDE(S)
Obj_opt = MAXVALUE;
for m = 1 to M do
    G = GMM(X,m);
    D = Discretize (X,G);
    Obj = Object (X,D,G);
    if Obj < Obj_opt then
      Obj_opt = Obj;
      D_opt = D;
return D_opt;
SubFunction Discretize (S, G)
for each s_i ∈ s do
    d = 1
    for j = 2 to 中心点的数量 do
      if P(s_i | G_j) > P(s_i | G_d) then
        d = j;
      D_i = d;
return D;
SubFunction Obeject(δ,D,G)
Obj = 0;
    for each G(μ_j,δ_j) ∈ G do
        Obj = Obj + 1n(2πeδ_j^2)/2
    for each di ∈ D do
        Obj = Obj + 0.5log(δ_di^2/δ^2) + (μ_i^2 + μ_di^2 + δ^2 - 2μ_i u_di)/(2δ_di^2)
```

1.2 近似解法

近似解在算法 1 中的应用如图 1 所示,主循环过程中,为了寻找最佳的离散化模型,此算法枚举了分量的所有可能数量。已知间距的数量,求目标函数值的过程分为三步: 首先构建一个标准的高斯混合模型,然后通过分配每个样本到拥有最高概率的分量中,从而使样本离散化,最后,在函数对象中计算出目标值,当使用算法 1 计算分量的可能的数量时,返回最佳离散化结果。

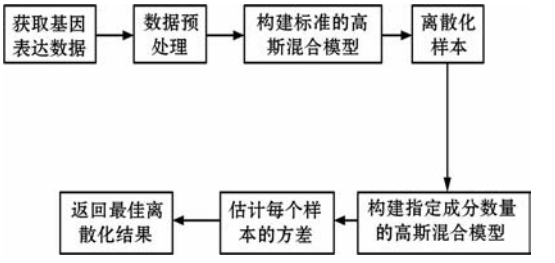

图 1 近似解法流程图

2 实 验

本节通过实验验证了所提的 RGMM 的有效性,算法在 Matlab 2010 下执行。所使用的电脑配置为:2.93GHZ DUO CPU、2GB RAM、1TB 硬盘。

2.1 数据集

实验使用了 6 个带有两个分类的分类数据集,包括:结肠癌[4]、弥漫大 B 细胞淋巴瘤(DLBCL)[7]、白血病[11]、肺癌[12]、前列腺癌[15]和 GCM[16]。

结肠癌数据集包括了 62 个患者样本,这些患者有的患有结肠癌,有的结肠组织正常。在 DLBCL 实验中,将样本分成两个子类的肿瘤,包括弥漫大 B 细胞淋巴瘤(DLBCL)和滤泡型淋巴瘤(FL)。白血病的数据集有两个分类组成,包括急性淋巴细胞白血病(ALL)和急性髓性白血病(AML)。实验中肺癌数据集包括两个子分类,MPM 和 ADCA。前列腺的实验的目标是把在正常的患者中选出患有前列腺癌的样本后再加以分类,而 GCM 实验的目标是把在正常的患者中选出患有成熟恶性肿瘤的样本后再加以分类。

2.2 实验过程及参数设置

RGMM 在离散化方法中的三种状态下比较。包括无监督的基于离散化方法的核密度估计法(KDE)[14],监督的离散化方法 SDM[5]、Khipos[5]、FUSINTER[6] 和 CAIM[7]。通过 KDE,SDM、FUSINTER 和 CAIM 得到的结果基于所提方法的算法执行,通过 Khipos 得到的结果基于 Khipos 系统。

在 UBR-LBR 关联规则挖掘框架里,一共有三个重要的准则,这些准则用来比较离散化方法、间距的数量、UBRs 的长度,以及分类精度。间距的数量和 UBRs 的长度是离散化数据和发现规则的复杂度的评价标准,分类精度反映出了发现规则的有用性。多重的符号测试[14]应用于此研究,其有效值级别 $\alpha = 0.1$。在下列实验中,在六个数据集中比较六个算法,此差别的临界值为 0,见文献[14]中的表 A.1。

2.3 实验结果及比较分析

1) 离散间距的数量

不同方法中离散化间距的平均数量见表 1,最后一排是 RGMM 获得量(只有很小的间距数量)减去 RGMM 丢失量之后的数值。根据多重符号测试,RGMM 的间距数量明显小于无监督方法 KDE 的间距数量,大于 SDM 和 CAIM 的间距数量,而且与 FUSINTER 的间距数量差不多大,它显示出了 RGMM 中使用规则因数的效果。离散化间距的数量也显示出最小描述长度很好地权衡了数据丢失和高斯混合模型中的熵。

表 1　不同离散化方法中的间距数量

数据集	无监督方法		监督方法			
	RGMM	KDE	Khiops	SDM	FUSINTER	CAIM
结肠癌	1.98	5.23	3.42	1.04	2.07	2.31
白血病	2.32	6.34	2.23	1.17	1.54	1.92
DLBCL	2.67	5.72	2.84	1.13	1.94	2.36
肺癌	2.08	4.62	3.18	1.11	2.38	2.54
前列腺癌	2.92	8.56	4.31	1.23	2.02	2.06
GCM	2.79	9.01	6.23	1.84	2.82	3.10
获得-丢失	—	-6	-4	6	0	-2

2) UBR 规则中的平均长度

表 2 中的最后一排显示了 RGMM 获得量(只有很短的 UBR 长度)减去 RGMM 丢失量之后的数值。如表 2 所示,无监督方法(RGMM 和 Khiops)的 UBRs 平均长度明显小于使用监督方法(Khiops、SDM、FUSINTER 和 CAIM)UBRs 长度,验证了所提方法的观察结果,监督离散化方法一定程度上引起了基因表达规则挖掘的联合展开现象。

表 2　UBR 规则的平均长度

数据集	无监督方法		监督方法			
	RGMM	KDE	Khiops	SDM	FUSINTER	CAIM
结肠癌	87	76	139	173	157	169
白血病	135	89	156	178	181	146
DLBCL	178	102	189	215	204	201
肺癌	123	113	152	225	205	184
前列腺癌	201	152	391	532	432	497
GCM	342	298	378	487	409	340
获得-丢失	—	6	-4	-6	-6	-4

比较两种无监督方法,RGMM 的 UBR 长度要比 KDE 的长得多,这是因为 KDE 把数据离散化成间距的数量要比 RGMM 多(见表 1 所示)。基于这种性质,KDE 对分类准确度有着不良的影响,关于这一点,可见表 3 所示。

表 3　平均分类准确度(%)

数据集	无监督方法		监督方法			
	RGMM	KDE	Khiops	SDM	FUSINTER	CAIM
结肠癌	92.12	85.12	87.23	90.32	90.32	91.59
白血病	94.32	83.79	92.32	95.28	94.32	95.28
DLBCL	88.55	81.35	85.68	86.88	85.01	84.35
肺癌	95.23	76.62	95.32	99.13	97.42	94.23
前列腺癌	80.01	70.11	79.32	77.28	74.23	71.93
GCM	85.41	75.05	82.21	94.15	84.42	83.28
获得-丢失	—	6	-4	0	-4	-4

对比这五种离散化方法,与 KDE 相比较,RGMM 可以获得相对较短的 UBR 规则,而且 RGMM 获得的 UBR 规则长度远远短于 Khiops、SDM、FUSINTER 和 CAIM 下的 UBR 规则长度,实验表明,RGMM 得到的离散化阈值对于数据标签而言并没有过拟合。

综上可得,RGMM 使用一种相对简单的模型(见表 1 和表 2 所示),可以得到一个很高的分类准确度(见表 3 所示),而且 RGMM 可以很好地权衡模型复杂度与经验风险之间的关系。

3　结　语

经过大量的实验发现,监督离散化方法可使离散化阈值过拟合于标签,这在一定程度上使基因表达规则挖掘中产生了规则展开现象。在最小化描述长度框架下,通过挖掘离散化模型复杂度及信息丢失准则,本文提出了基于离散方法的正规化高斯混合模型(RGMM)。应用 RGMM 获得的实验结果表明,无监督离散化模型不仅可以寻找到相应的离散化阈值,还可以显示出离散化在关联规则挖掘中规则展开现象问题中的重要性,这是以前研究中常常被忽视的问题。

未来会将所提方法运用在不同的数据集上,改变初始参数的设置,并且进一步改善分类准确度。

参　考　文　献

[1] 高阳. 中国数据挖掘研究进展 [J]. 南京大学学报:自然科学版,2011,47(4):351-353.

[2] Janez Demšar, Blaž Zupan. Orange: Data Mining Fruitful and Fun A Historical Perspective [J]. Informatica, 2013, 37(3): 55-60.

[3] Kurgan L, Cios K. Caim discretization algorithm [J]. IEEE Trans Knowl Data Eng, 2011, 16(2): 145-153.

[4] Luengo J, Saez J, Lopez V, et al. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning [J]. IEEE Trans Knowl Data Eng, 2012, 33(4): 109-122.

[5] 覃光华,李祚泳. BP 网络过拟合问题研究及应用 [J]. 武汉大学学报:工学版,2010,39(6):55-58.

[6] Biba M, Esposito F, Ferilli S, et al. Unsupervised discretization using kernel density estimation [J]. International joint conference on artificial intelligence, 2012, 34(5): 696-701.

[7] 孟祥福,张霄雁,马宗民,等. 一种基于领域知识的 XML 数据模糊查询方法 [J]. 智能系统学报,2012,7(6):50-53.

[8] Luengo J, Saez J, Lopez V, et al. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning [J]. IEEE Trans Knowl Data Eng, 2012, 35(2): 132-142.

[9] 李海军,王钲旋,王利民,等. 一种基于贝叶斯测度的有监督离散化方法 [J]. 仪器仪表学报,2011,26(8):786-789.

[10] Botev Z, Grotowski J, Kroese D. Kernel density estimationvia diffusion [J]. Ann Stat, 2010, 38(5): 2916-2957.

[11] 花海洋,赵怀慈. 一种新的无监督连续属性离散化方法 [J]. 计算机工程与应用,2011,47(6):208-211.

[12] Flores M, Gámez J, Martínez A, et al. Handling numeric attributes when comparing Bayesian network classifiers: does the discretization method matter [J]. Appl Intell, 2011, 34(2): 372-385.

[13] García S, Fernández A, Luengo J, et al. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power [J]. Inf Sci, 2010, 33(10): 2044-2064.

[14] Amorim R, Mirkin B. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering [J]. Pattern Recognition, 2012, 45(10): 61-75.

[15] Gupta A, Mehrotra K, Mohan C. A clustering-based discretization for supervised learning [J]. Stat Probab Lett, 2010, 80(9): 816-824.

[16] Spirduso W, Reeve T G. The national academy of Kinesiology 2010 review and evaluation of doctoral programs in Kinesiology [J]. Quest, 2011, 63(4): 411-440.