粗糙自组织映射在基因表达数据分析中的应用*)

杨 迪 苗夺谦 王睿智

(同济大学计算机科学与技术系 上海 201804)

(嵌入式系统与服务计算教育部重点实验室(同济大学) 上海 201804)

摘 要 本文利用粗糙集与布尔逻辑离散约简算法改进了粗糙自组织映射算法,并应用于基因表达数据的分析中。算法改进了传统自组织映射收敛慢、网络规模难以确定的缺点,减小了网络规模不确定对分类效果的影响。使用酵母菌基因表达数据进行实验,得到了较好的网络质量、网络规模和分类效果,相比传统自组织映射使分类正确率提高了10.15%。

关键词 Rough set,自组织映射网络,基因表达数据

Analyzing Gene Expression Data Based on Rough Self-organizing Maps

YANG Di MIAO Duo-Qian WANG Rui-Zhi

(Department of Computer Science and Technology, Tongji University, Shanghai 201804)

(Key Laboratory of Embedded Systems and Service Computing (Tongji University), Ministry of Education, Shanghai 201804)

Abstract This paper presents an approach using rough set and Boolean reasoning discretization-reduction to improve the rough self-organizing maps, and using it to analyzz gene expression data. This approach can improve the convergence time and obtain the best map scale. We have tested the RSOM on the yeast data set, obtained better map quality and classification result.

Keywords Rough set, SOM, Gene expression data

1 引言

基因芯片技术的发展使得我们可以方便地获得海量基因表达数据。如何有效地分析这些海量数据并从中获得有价值的信息变成当今生物学和信息科学的热点话题。目前已经有了许多有效的途径,比如:K 聚类,神经网络,支持向量机等。

自组织映射网络是一种无监督的机器学习方法。它能把高维数据表示成低维(通常是 2 维)数据,可方便地完成复杂数据的可视化,并易于解释结果,对微小扰动有较好的鲁棒性。Tamayo. P (1999)[1]曾利用自组织映射网络分析基因表达数据。之后还有学者不断尝试应用自组织映射网络在这方面的应用如 Sugiyama A(2002) [2]使用 K-聚类与 SOM 结合,Shigeru Takasaki (2006) [3]应用自组织映射选择有效 siRNA 序列。但是传统自组织映射网络收敛速度慢,且局部自振荡;网络连接权的初始状态、神经元个数等参数选择对网络的收敛性和分类效果有较大影响。Sanker(2004)[4]提出粗糙自组织映射网络的概念,利用粗糙集理论解决自组织映射网络的上述局限。鉴于自组织映射网络过去在基因表达数据上的广泛应用和成果,本文改进了粗糙自组织映射算法,并应用于基因表达数据分析。

基因表达数据的特点是:属性比较多、属性冗余度高、对象比较少、主要为连续数据,这使得传统自组织映射网络的许多缺陷暴露出来,分类效果也受到影响。因此针对基因表达数据考虑首先对数据进行属性离散化和约简,去除对分类没有贡献的冗余属性,利用约简结果设置网络初始连接权值和神经元规模。Sanker 模糊粗糙自组织映射方法是使用模糊集方法对数据离散化,之后用差别矩阵方法对数据集约简。本文使用粗糙集与布尔代数的离散约简算法,改进粗糙自组织映射网络,既克服了传统自组织映射的缺点,使预估有效神经元规模成为可能,减小了随机网络规模对分类效果的影响,又通过离散和约简同步进行的思想提高了算法的效率。相对于 Sankar 模糊粗糙自组织映射模型,优化了时间、空间复杂度和发生误差的概率。实验结果表明,使用此算法分析酵母菌基因表达数据,得到较好结果。

本文主要包括 5 部分:第 1 部分,引言介绍研究背景和进展;第 2 部分介绍了粗糙集与布尔逻辑结合的离散约简算法;第 3 部分主要介绍改进的粗糙自组织映射算法;第 4 部分为基因表达数据的应用和结果分析;最后是结论。

2 粗糙集与布尔逻辑结合的离散化规则生成方法

下面介绍 Nguyen 和 Skowron 提出的启发式粗糙集与布尔逻辑离散约简算法[5]。

算法 1 启发式粗糙集与布尔逻辑离散约简算法首先构造一个信息表 $CUT=\emptyset$ 如下:

对于任意 P_r^a 为属性 a 的第 r 个断点,如果

$$[c_r^a, c_{r+1}^a) \subseteq [\min(a(x_i), a(x_j)), \max(a(x_i), a(x_j)))$$

那么 $P_r^a((x_i,x_j))=1$;否则 $P_r^a((x_i,x_j))=0$。

输入:SPR:单一决策区集合;
SIR:多决策区集合;
S:为原决策表按每个属性 $P_r^a((x_i,x_j))$ 改写的信息表。

输出:断点集 CUT,规则集 SPR。

*)国家自然科学基金项目(60475019)博士学科点专项科研基金(20060247039)。**杨 迪** 硕士研究生,主要研究方向:模式识别与智能系统、粗糙集理论、数据挖掘;**苗夺谦** 教授,博士生导师,主要研究方向:人工智能、模式识别、知识发现、粗糙集理论等。

启发式算法：
①初始化状态：根据原来的信息表 S 构造新的信息表 S^*，令 $SPR=\varnothing$，$SIR=\{\mathcal{R}^*\}$，初始化断点集 $CUT=\varnothing$；
② 选取所有列中 1 的个数最多的断点加到 CUT 中，去掉此断点所在的列和在此断点上值为 1 的行；当有一个以上断点的列 1 的个数相同时，把列对应的断点所在的列值为 1 的行的 1 的数目相加，取和最小的断点。
③if ∀ Re g∈SIR 继续循环
if CUT 把 Re g 分为两个区域 Re g1，Re g2 则
{把 Re g 从 SIR 中去掉；
if Re g1 中所有对象都有相同决策值 v_d
then 把规则[(x∈Re g1)⇒$d(x)=v_d$]插入到 SPR 中
else 把 Re g1 插入到 SIR
if Re g2 中所有对象都有相同决策值 v_d
then 把规则[(x∈Re g2)⇒$d(x)=v_d$]插入到 SPR 中
else 把 Re g1 插入到 SIR }
④ 如果信息表 S^* 中元素不为空，则转到②；否则停止。此时 CUT 即是得到的断点集。

3 粗糙自组织映射算法

3.1 自组织映射网络

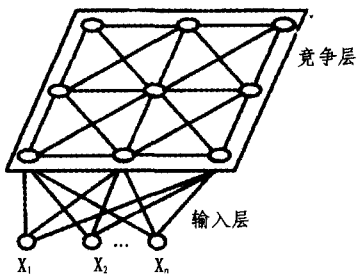

图 1 SOM 结构图

Kohonen 的 SOM 网络[6]是一单层向前网络，网络结构相当简单(如图 1 所示)，SOM 网络只有输入层和输出层，而且两层之间为完全连接。

设输入矢量 g 的维数为 n 维，则输入层有 n 个节点；输出层由 m 个神经元组成的二维平面阵列；输入层与输出层各神经元之间完全连接，设连接权重向量为 w_{ij}，其中 i 表示输入层的第 i 个节点，j 表示输出层的第 j 个节点。输入层节点并不对输出矢量 x 进行处理，它只是把输入矢量的各个分量经过连接权传送到输出层，激发输出层各个神经元之间的竞争，最后输出获胜神经元，因此输出层也称竞争层。

3.2 粗糙自组织映射算法

Sanker 模糊粗糙自组织理论采用粗糙集约简后的规则初始化自组织映射的连接权值和神经元个数，避免了使用随机权值训练引起的收敛速度慢、局部自振荡等缺陷，加快收敛速度；并设置自组织映射的神经元个数为约简后规则中不同析取范式个数，解决了网络连接权的初始值、神经元个数等参数不确定影响分类效果的问题。

但是 Sanker 使用模糊集离散化，差别矩阵约简的方法相对于大规模、高冗余的基因数据有着一定局限性。此种方法使所有属性都被离散化为高、中、低三部分，离散化标准太过于固定，许多对分类支持度不高的属性仍然继续参与约简。此外模糊集方法离散化中属性一般很少被约简，数据的海量和高冗余又转移到差别矩阵约简阶段，使得分类效果受到影响。

因此本文考虑使用布尔逻辑与粗糙集离散约简算法改进 Sanker 算法。布尔逻辑与粗糙集离散约简过程实际上是在保持信息系统的不分明关系的前提下寻找最小断点集和必要规则集，所以在去除冗余属性的同时保持了训练数据的分类能力。

算法 2 改进的粗糙自组织映射算法

输入：训练集决策表 S；
输出：完成训练的自组织映射网络。
① 使用布尔逻辑与粗糙集结合的离散化方法对数据集进行离散化，得到断点集 CUT 和规则集 SPR。
② 由规则集 SPR 得
$R\equiv[(l_1\leqslant a_1<r_1)\wedge\cdots\wedge(l_k\leqslant a_k<r_k)\Rightarrow(d=v_d)](R\in SPR)$；
由断点集 CUT，改写决策表 $S\Rightarrow S'$；
参考 S' 转化规则为：
$R^P\equiv[(a_1^P=v_{1,1})\vee(a_1^P=v_{1,2})\vee\cdots]\wedge\cdots\wedge[(a_k^P=v_{k,1})\vee\cdots]\Rightarrow(d=v_d)$
转化合取范式为析取范式：
$O_i^P\equiv[(a_1^P=v_{d1,1})\wedge(a_2^P=v_{d1,2})\wedge\cdots]\vee\cdots\vee[(a_1^P=v_{dN,1})\wedge(a_2^P=v_{dN,2})\wedge\cdots]$
O_i^P 为第 i 种决策的对象集，不同析取范式个数为 N_i。
③ 确定自组织映射的神经元个数：
$\because S$ 在分类能力上等价于 S'，$N=\sum N_i$ 为 S' 中不同的析取范式；
$\therefore$ 自组织映射网络神经元个数为 N。
④ 初始化自组织映射权值表：
对于每个属性 i 的每种输入
if$(a_i^P=v_{dn,i})\in O_i^P$
$w_{dn,i}:=High$
else
$w_{dn,i}=Low$
⑤ 利用离散化后的数据集训练自组织映射。

4 实验数据分析

4.1 实验数据

本文采用的数据是斯坦福大学分子生物系网站上公布的六千多条酵母细胞有丝分裂过程中起调控作用的基因表达数据[7]。在生物上酵母(Yeast)是人类已经测出基因全序列的真核生物之一，并且酵母中的许多基本遗传机制是与人类相似的。因此，选用该数据不仅有生物学上的意义，也有医学上的意义。选择其中经过规范化的 237 组酵母菌细胞周期每组为 17 维，可分为 4 类具体数据如表 1。

表 1 实验数据说明

类别	描述	个数
1	G1	49
2	S	31
3	G2	18
4	M	139

4.2 评价标准

下面介绍两种对于自组织映射质量的评价标准。

4.2.1 量化误差

量化误差 q_E 衡量训练时竞争层中获胜神经元的权值向量与输入向量拟和的速度。公式如下：

$$q_E=\frac{\sum_{p=1}^{n}(\sum_{all\_wining\_node}\sqrt{(\sum j(x_{pj}-m)^2)})}{number\_of\_patterns}$$

这里 $j=1,\cdots,m$，m 是输入属性的个数，x_{pj} 是第 p 种模式的第 j 个成员，共有 n 种模式。量化误差越高权值向量与输入向量的差别越大。

4.2.2 拓扑误差[8]

为了衡量映射图的连续性，引入拓扑误差 ε_t。假设一个样本向量 $x\in M$，最靠近样本的向量为 w_i，第二靠近的为 w_j。假设 n_j 中在 x 和 w_j 之间的点都映射到 w_i，其他的点映射到 w_j。如果相对应的神经元 n_i 和 n_j 相邻，映射为局部连续的；如果它们不相邻，则存在局部的不连续，或者拓扑误差。对于整个映射的拓扑误差 ε_t 是对所有样本向量中的局部拓扑误差进行求和然后标准化：

$$\varepsilon_t=\frac{1}{n}\sum_{k=1}^{N}u(x_k),\text{where}$$

$$u(x_k)=\begin{cases}1,\text{最优或次优匹配不邻近的单元}\\0,\text{其他}\end{cases}$$

这样定义，ε_t 表明了局部相邻关系的正确性。

4.3 实验结果与分析

本文是基于芬兰赫尔辛基大学开发的 SOM Toolbox 工具箱[9]中的标准程序进行开发和改进的。实验数据从基因表到数据中随机抽取 80%和 20%，70%和 30%，51%和 49%，分别作为训练集和测试集(训练集和测试不相交)，分别进行 5 次，3 次，1 次实验。之后实验过程分为两部分，一部分不经离散化直接使用自组织映射训练；另一部分使用粗糙自组织映射算法训练。实验结果如表 2。

表 2 实验结果

实验类型	网络规模	识别率	量化误差	拓扑误差
190_47SOM	24 * 3	79.44%	0.95846	0.05146
166_71SOM	21 * 3	81.06%	0.9486	0.07033
120_117 SOM	8 * 7	78.02%	0.9829	0.2203
190_47 RSOM	14 * 5	86.42%	0.78736	0.01478
190_47 RSOMT	18 * 10	89.67%	0.51912	0.00422
166_71 RSOM	11 * 10	88.33%	0.80903	0.01033
166_71 RSOMT	16 * 8	91.77%	0.51067	0.00400
120_117 RSOM	14 * 4	84.89%	0.9626	0.0254
120_117 RSOMT	12 * 8	87.52%	0.6369	0.0109

"190_47"表示训练集测试集数量比，SOM 表示自组织映射，RSOM 表示粗糙自组织映射，RSOMT 表示确定神经元个数的自组织映射。

各种实验样本结果比较图如下。

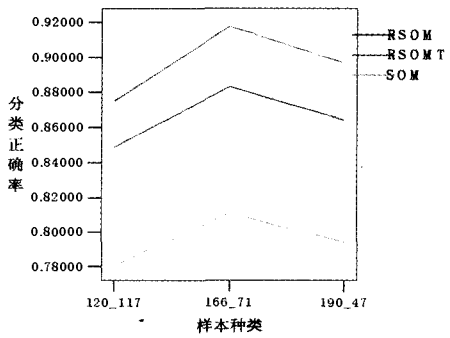

图 2 分类正确率图

通过表 2 发现粗糙自组织网络无论是在识别率，还是量化误差和拓扑误差上都使映射网络性能有了显著的提高。利用粗糙自组织映射约简后确定的神经元个数代入训练参数后确定的映射网络质量和分类效果都有明显提高。

RSOM 相比 SOM 在分类能力上差别明显如图 2 所示；在量化误差上 RSOM 相对 SOM 对网络规模变化影响较大如图 3 所示；RSOM 对比 SOM 在拓扑误差方面表现出对于网络规模变化的更小的波动性、更稳定的网络结构。

结论 粗糙集理论能够对信息系统进行约简，从而产生规则，但规则比较零乱、可视性不强，不易理解；自组织映射收敛比较慢、神经元个数难以确定。本文使用改进的粗糙自组织映射算法应用于基因表达数据分析，结合了两者的优点，有效地提高了分类效果。同时利用布尔逻辑与粗糙集结合的理论使离散化和约简同时进行，简化了粗糙自组织映射算法步骤，减小了误差发生率。

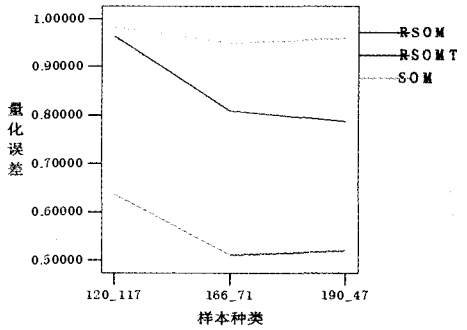

图 3 量化误差

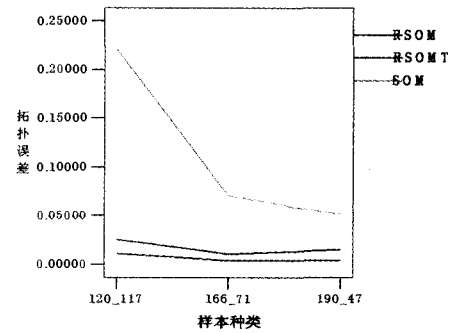

图 4 拓扑误差

粗糙自组映射方法是有效基因表达数据分析的工具之一，能够处理属性数较多并且高冗余的数据集，能够广泛用于数据分类和展示。但在数据量限制和进一步提高分类效果上还有进一步研究的空间。结合粒度计算[10]的思想，粗糙自组织映射理论在可解释性和算法适用范围上将有更大的发展。

参 考 文 献

1 Tamayo P, et al. Interpreting Patterns of Gene Expression With Self-organizing Maps: Methods and Application to Hematopoietic Differentiation. In: Proc. Natl. Acad. Sci. USA, 1999, 96: 2907～2912

2 Sugiyama A, Kotani M. Analysis of gene expression data by using self-organizing maps and k-means clustering. In: Neural Networks, 2002. IJCNN '02. 2002, 2: 1342～1345

3 Takasaki S, et al. Selecting effective siRNA sequences based on the self-organizing map and statistical techniques. In: Computational Biology and Chemistry, 2006, 30: 169～178

4 Sanker K PL, et al. Rough Self Organizing Map. In: Applied Intelligence, 2004, 21: 289～299

5 Nguyen H S, Skowron A. Quantization of real values attributes, rough set and Boolean reasoning approaches. In: JACIS'95, 1995. 34～37

6 Kohonen T. The self-organizing map. In: Neuracomputing, 1998 (21): 1～6

7 Kimmo Kiviluoto. Topology Preservation in Self-Organizing Maps. In: PICNN'96, 1996. 294～299

8 Yeung K Y, Ruzzo W L. Principal component analysis for clustering gene expression data. In: Bioinformatics, 2001, 17(9): 763～774

9 Vesanto J, et al. SOM Toolbox for matlab. http://www.cis.hut.fi/projects/somtoolbox

10 李道国，苗夺谦，杜伟林. 粒度计算在人工神经网络中的应用. 同济大学学报，2006，7：960