基于多维自组织特征映射的聚类算法研究*)

江 波 张 黎
(贺州学院计算机科学与工程系 贺州 542800)

摘 要 作为神经网络的一种方法,自组织特征映射在数据挖掘、模式分类和机器学习中得到了广泛应用。本文详细讨论了自组织特征映射的聚类算法的工作原理和具体实现算法。通过系统仿真实验分析,SOFMF 算法很好地克服了许多聚类算法存在的问题,在时间复杂度上具有良好的性能。
关键词 组织特征映射,聚类,数据挖掘,神经网络

Study of Algorithms of Clustering Based on Multi-dimensional Self-organizing Feature Mapping

JIANG Bo ZHANG Li
(Department of Computer Science and Engineering, Hezhou University, Hezhou542800, China)

Abstract As a method of neural network, the self-organizing feature mapping(SOFM) is an excellent approach for data mining, pattern classification and machine learning. The theory and algorithm of SOFM are discussed in detail in this article. Simultaneously analyze and summarize this algorithm; overcome the insufficiency of many clustering algorithms, be able to find clusters in different shapes, be non-sensitive to the input data sequence, process noise data and multi-dimensional data well, and have multi-resolution.
Keywords Self-organizing feature mapping, Clustering, Data mining, Neural network

1 引言

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别人们事先不知道的、有效的、新颖的、潜在有用的,以及最终可理解的知识的过程[1-4]。它是一门涉及面很广的交叉学科,包括神经网络、模式识别、机器学习、数据库、数理统计、粗糙集、模糊数学等相关技术,是知识发现(Knowledge Discovery in Database)的关键步骤。数据挖掘意味着在大量事实或观察数据的集合中发掘信息和知识的决策支持过程,提取的知识可以表示为概念(Concepts)、归则(Rules)、规律(Regularities)、模式(Patterns)等形式[2-4]。

在数据挖掘的算法中,聚类技术已成为信息处理的核心技术。从 20 世纪 40 年代至今,国内外的研究者提出了很多聚类算法,如基于层次的、基于平面分割、基于密度、基于规则和基于网格的算法等,在这些众多的算法中,大多数算法都需要事先人为地给出一些参数,而且时空效率也有待于进一步提高。然而,在没有先验知识的情况下,人为地确定这些参数是十分困难的。为了解决这个难题,需要研发新的聚类算法,在保证不降低时空效率和信息处理性能的前提下,力图减少或避免需要事先人为确定的参数。

自组织特征映射(Self-Organizing Feature Mapping, SOFM)[1]神经网络作为聚类技术的一种,是由芬兰赫尔辛基大学神经网络专家 Kohonen 教授在 1981 年提出的,是一种聚类方法,它能根据其学习规则对输入的模式自动进行分类,即在无监督的情况下,对输入模式进行自组织学习,通过反复地调整连接着输入和输出的权重系数,最终使得这些系数反映出输入样本之间的相互距离关系,并在竞争层中将分类结果表示出来。因此 SOFM 网络在结构上模拟了大脑皮层中神经元呈二维空间点阵的结构,在功能上通过网络中神经元间的交互作用和相互竞争,模拟了大脑信息处理的聚类功能、自组织和自学习功能[1]。

SOFM 网络由输入层和输出竞争层组成,如图 1 所示,两层之间是全互连的。SOFM 能将任意输入模式在输出层进行反映,并保持其拓扑结构不变。在输出竞争层,获胜的神经元邻域内的神经元在不同程度上都得到兴奋,随着时间增大,邻域减少,最后可能剩下一个或者一组神经元,最终得到的区域反映了一类输入模式的属性[5,6]。SOFM 根据学习规则,通过对输入模式的自组织,能在竞争层将分类结果表示出来。它不是用一个神经元来反映分类结果,而是以若干神经元同时反映。通过对输入模式的反复学习,可以使连接权矢量空间的分布密度与输入模式的概率分布趋于一致,尽可能地反映输入模式的统计特征。

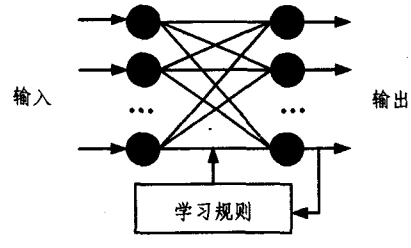


图 1 SOFM 网络模型

自组织输出映射的相邻神经元可以识别输入空间的邻域,SOFM 能够识别训练输入向量的分布和拓扑结构[1]。

2 SOFM 算法基本原理

自组织特征映射通过 D 里的样本 X 学习拓扑映射 $f:D\subset R^s\to G\subset R^m$,其中 G 是包含一组神经元的输出映射,每个神经元代表 m 维欧几里德空间的一个元素。$r_i\in G$ 表示输出映射的第 i 个神经元(i^{th})的位置。$X=[x_1,x_2,\cdots,x_s]\in D$ 是输入向量。假设每个输入向量都并行连接到输出映射的每个神经元。神经元 i 的权重向量表示为 $W_i=[w_{i1},w_{i2},\cdots,w_{is}]^T\in R^s$[1]。根据学习规则:

$W_i(t+1)=W_i(t)+\alpha(t)\lambda(i,i^*)[X(t)-W_i(t)]$,其中,$t=1,2,3,\cdots$是离散时间坐标系,$\alpha(t)=1,2,3,\cdots$是学习速率因子,$\lambda(i,i^*)$是邻域函数。获胜神经元 i^* 定义为权重向量在输入空间 $X(t)$有最小的欧几里德距离的神经元:

$\|W_{i^*}(t)-X(t)\|\leqslant\|W_i(t)-X(t)\|,\forall r_i\in G$。

*)重庆市科委自然科学基金计划资助项目(No. CSTC 2007BB2451)。江 波 讲师,主要研究方向为数据挖掘、信息处理等;张 黎 讲师,主要研究方向为数据挖掘、数字化校园、信息处理等。

Kohonen 提出一个典型的邻域函数：

$$\lambda(i,i^*)=\begin{cases}1 & \text{for} \parallel r_i-r_{i^*} \parallel \leqslant N_{i^*}(t)\\ 0 & \text{otherwise}\end{cases}$$

其中 $N_{i^*}(t)$是某个离散时间函数。

利用 SOFM 聚类时，输入模式的规模不一，维数有高有低，以及蕴涵的聚类个数、各个聚类的大小、形状，以及分布密度都是未知的，甚至有些模式点还是噪音之类。

SOFM 工作原理[1]如下：

假设网络有 M 个输入节点，N 个输出节点，基本的 SOFM 的工作原理如下：

(1) 初始化输入样本矢量集 $x=(x_1,x_2,\cdots,x_M)$

(2) 取训练样本矢量，按照一定的失真测度计算输入矢量和输出节点之间的失真，选择最小失真的输出节点作为获胜节点：

$$d_j=\sum_{i=1}^{M}(x_i(t)-W_{ji}(t))^2,\ 1\leqslant j\leqslant N$$

$$d_j^*=\min_{1\leqslant j\leqslant N}\{d_j\}$$

(3) 调整获胜节点及其邻域内节点的权值：

$$w(t+1)=w(t)+\alpha(t)(x(t)-w(t))$$

(4) 提供下一模式，直到全部模式学习完毕。更新学习速率和邻域函数，进入(2)重新学习，直到算法收敛。

3 SOFM 算法实现

(1) 初始化权重系数，$[W_{ji}]$赋予[0,1]区间的随机值，选择邻域半径 r，以及学习速率 $\eta(t)$；

(2) 对于每个输出神经元计算欧氏(Euclidean)距离 d_j，找出最小的欧氏距离 d_g，并确定获胜神经元 g；

(3) 对处于获胜神经元邻域内的所有神经元，调整它们的权重系数：$W_{jk,i}^{(t+1)}=W_{jk,i}^{t}+\eta(x_i-W_{jk,i}^{t})$，位于邻域之外的神经元不作调整，且邻域不应当包括边界；

(4) 更新学习速率和邻域半径(不断减小)；

(5) 重复上述第(2)到第(4)步，直至满足条件，则停止。

4 SOFMF 算法

利用自组织特征映射网络族(SOFMF)进行聚类的思想来自于团队思想，SOFMF 利用一族自组织特征映射网络来工作。虽然，单个的自组织特征映射网络有缺点，但是，自组织特征映射网络族里的每个成员可以一起工作，互相纠正，从宏观上来讲，可以得到一个令人满意的聚类算法。

首先构造一族具有不同大小的输出映射的自组织特征映射网络。图 2 所示的是一个具有二维输出映射的 SOFMF。随着成员索引号的增加，输出映射的大小缓慢地、单调地增加。SOFMF 的神经元不一定必须以二维的方式进行管理，还可以是一维或者是三维，甚至更高维。这样，就可以用类似的方式构造出一个 SOFMF。

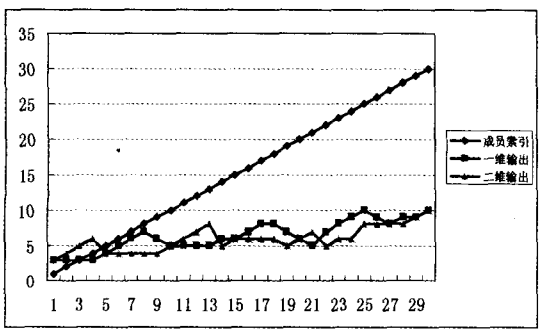


图 2 一个具有二维输出映射的 SOFM

使用输入数据集 P 把 $SOFM_k$ 训练若干步，这样就可以使用 P 对 $SOFM_k$ 进行仿真，得到 P 的 c_k 划分。最后的结果表明，如果以欧几里德距离作为相似性来度量的话，在同一个获胜区里的对象有更高的相似性。由于 P 的 c_k 划分是一种具有拓扑保持性的划分，因此，可以采用拓扑相似性(topological similarity)来定义各对象之间的相似性[1]。

假设 $X(p),X(q)\in P$，则相似性为：

$$TS_k(X(p),X(q))=\begin{cases}1 & ,X(p)\in N_k(i)\text{ and }X(q)\in N_k(i)\text{ for some }i\in G_k\\ 0 & ,X(p)\in N_k(i)\text{and }X(q)\in N_k(j),i\neq j,i,j\in G_k\end{cases}$$

如果在一个下三角形矩阵里存储 n 个对象每对之间的相似性，则可以得到一个拓扑相似性矩阵，用 TSM_k 表示：

$$TSM_k=\begin{bmatrix}1 & & & & \\ TS_k(X(2),X(1)) & 1 & & & \\ TS_k(X(3),X(1)) & TS_k(X(3),X(2)) & 1 & & \\ \vdots & \vdots & \cdots & \cdots & \\ TS_k(X(n),X(1)) & TS_k(X(n),X(2)) & \cdots & \cdots & 1\end{bmatrix}$$

当 $c_k=1$ 时，TSM_k 是一个 $n\times n$ 的单位下三角矩阵；当 $c_k=n$ 时，TSM_k 是一个 $n\times n$ 的单位矩阵。

5 仿真实验结果分析

本实验随机生成 2000 个二维向量，作为样本，采用二维映射网络的神经元结构为 8×8，用不同的步长，训练网络，得到的仿真结果如图 3-5 所示。

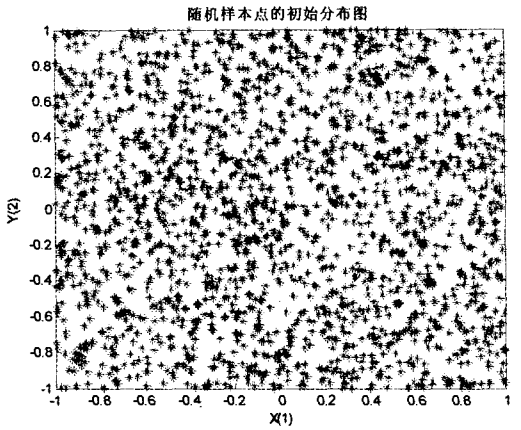


图 3 随机样本的初始分布图

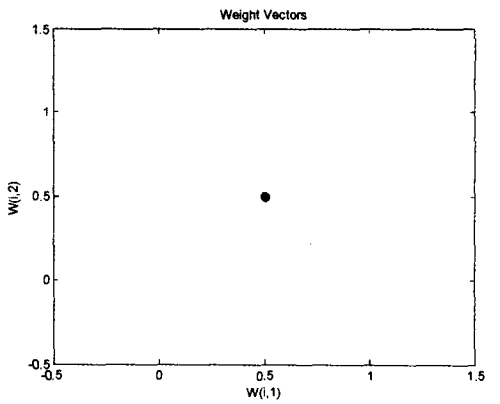


图 4 初始权值分布图

SOFM 族的每个成员都可以产生一个拓扑相似性矩阵，但是仅由单个的拓扑相似性矩阵几乎不能得到一个令人满意的聚类结果。

(下转第 185 页)

表 1 对训练样本进行识别时四种算法获得的正确率的比较

Data sets	BPNN	K-NN	SVM	FMSCA_MF
Australian	0.8184	0.7633	0.9687	1.0000
Balance	0.9260	0.9040	0.9753	1.0000
Diabetes	0.8923	0.8282	0.9381	1.0000
SVMguide2	0.9977	0.8553	0.8529	1.0000

表 2 对整个数据集中样本进行识别时四种算法获得的正确率的比较

Data sets	BPNN	K-NN	SVM	FMSCA_MF
Australian	0.7313	0.7173	0.8182	0.8425
Balance	0.9006	0.8976	0.9516	0.9530
Diabetes	0.8245	0.7962	0.8544	0.8906
SVMguide2	0.8269	0.8188	0.8309	0.9023

表 3 三种方法分类正确率的比较

Data sets	FMSCA_MF 的分类正确率	FMSCA _RS 的平均分类正确率	集成方法的分类正确率
Australian	0.5942	0.5913	0.5942
Balance	0.8800	0.8720	0.8800
Diabetes	0.7296	0.7264	0.7427
SVMguide2	0.7179	0.7391	0.7628

其次，在每个真实数据集上，我们抽取 60%的样本作为训练样本，余下的 40%作为测试样本。在每个真实数据集上我们利用随机选择球心的方法来执行新的模糊多球分类算法(Fuzzy Multi-spheres Classification Algorithm based the Random Selection，即 FMSCA_RS)，总共执行 10 次。我们将 10 执行后集成的投票结果、FMSCA_RS10 次执行的平均结果以及在同样的数据集上使用 FMSCA_MF 算法的结果在表 3 中进行了比较，其中较高的分类正确率已用粗体表示出来。从表 3 可以看出本文提出的集成方法其分类性能好于 FMSCA_MF 和 FMSCA_RS 的平均表现。

结束语 本文提出了一种新的模糊多球分类算法，该算法在训练阶段为每一个模式类构造多个球覆盖其所有的训练样本，并且在识别阶段利用一个模糊隶属函数来归类测试样本。此外，在提出的分类算法的基础上，还给出了它的集成方法。最后，在四个真实数据集上进行实验，实验的结果表明本文提出的算法及其集成方法不仅可以识别全部训练样本，而且具有较好的分类性能。

参 考 文 献

[1] Kulkarni S R, Lugosi G, Venkatesh S S. Learning pattern classification-a survey[J]. IEEE Trans. Inf. Theory, 1998, 6(44): 2178-2201

[2] Jain A K, Duin R P W, Mao J. Statistical pattern recognition: a review[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2000, 1(22): 4-33

[3] Dasatrthy B V. Nearest neighbor(NN) norms: NN pattern classification techniques[M]. Los Alamitos, CA: IEEE Computer Society Press, 1990

[4] Qiang G Q, Zhang P. Neural networks for classification: a survey[J]. IEEE Trans. Syst. Man Cybern. Part C, 2000, 4(30): 451-458

[5] Burges J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, 2: 121-167

[6] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge, U. K: Cambridge Univ. Press, 2000

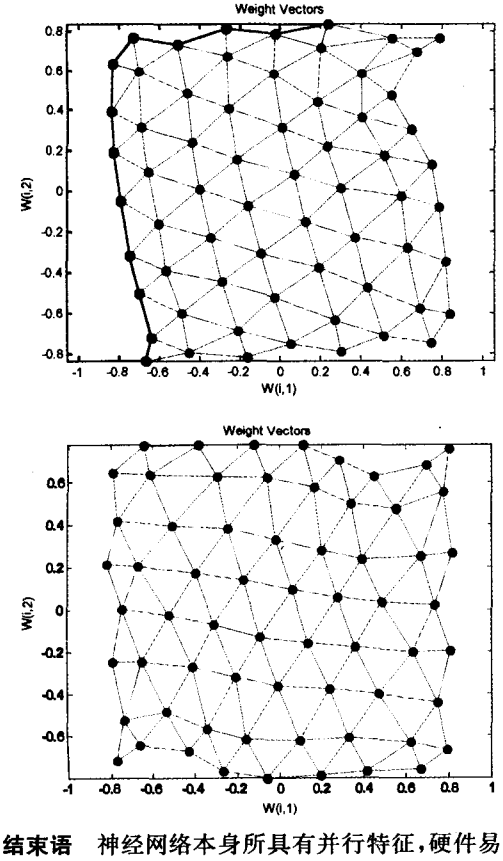
[7] Yager P R, Filev D P. Approximate clustering via the mountain method[J]. IEEE Trans. Syst. Man Cybern., 1994, 8(24): 1279-1284

[8] UCI Machine Learning Repository[DB]. http://www.ics.uci.edu/-mlearn/MLRepository.html

[9] Statlog coolection[DB]. http://www.niaad.liacc.up.pt/old/statlog/datasets.html

[10] LibSVM Website[DB]. http://www.csie.ntu.edu.tw/-cjlin/libsvmtools/dataset

(上接第 182 页)



结束语 神经网络本身所具有并行特征，硬件易实现性等，SOFMF 算法很好地克服了许多聚类算法存在的问题，在时间复杂度上具有良好的性能。针对不同的用户，可以得到不同层次的聚类结果，以便详细地进行分析。

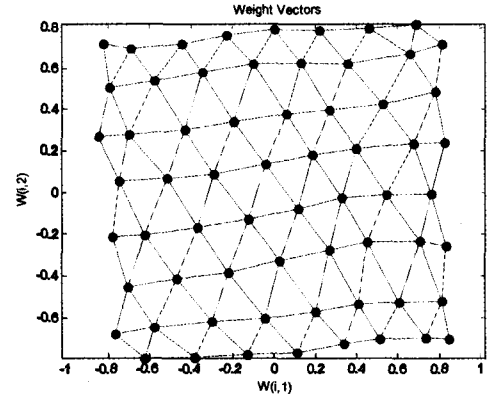


图 5 针对不同的步长，训练网络，相应的权值分布图

参 考 文 献

[1] Ezequied L R, et al. Invariant pattern identification by self-organising networks. Pattern Recognition Letters, 2001, 22: 983-990

[2] Dunkel B, Soparkar N. Data Organization and Access for Efficient Data Mining. ICDE, 1999

[3] Han J, Fu Yongjian. Mining Multiple-Level Association Rules in Large Database. IEEE Trans. on Knowledge and Data Engineering, 1999, 11(5): 798-805

[4] 刘君强，孙晓莹，潘云鹤. 关联规则挖掘技术研究的新进展. 计算机科学，2004，31(1)：110-113

[5] Bloch, Isabelle. Fuzzy relative position between objects in image processing: A morphological approach. IEEE Transactions on Patten Analysis and Machine Intelligence, 1999, 21(7): 657-664

[6] Brin S, Motwai R J D, Ullman, et al. Dynamic Itemset Counting and Implication Rules for Market Basket Data // ACM SIGMOD Conference on Management of Data. 1997: 265-276