一种生长型自组织神经网络的聚类研究

傅 雪,张少白
(南京邮电大学 计算机学院,江苏 南京 210003)

摘 要:自组织特征映射神经网络SOM(Self-Organizing Feature Maps)是一种优良的聚类工具,但其存在着一些限制,如需要预先定义网络大小、网络的收敛性较差和结构不灵活等。为了克服这些不足,在自组织神经网络理论的指导下,提出了一种基于生长型自组织神经网络的聚类方法。在无监督的情况下,该方法采用阈值控制的触发机制实现网络中神经元的生长和删除,并通过神经元权值的有效调整,以期得到数据对象的聚类结果。实验以二维空间中的数据对象为输入样本,验证了该方法的有效性和优越性。

关键词:自组织;生长;特征映射;聚类;神经网络
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2011)03-0064-03

Clustering Study of a Growing Self-Organizing Neural Network

FU Xue, ZHANG Shao-bai
(Computer College, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

Abstract: The self-organizing feature maps is a good clustering tool, but there are some restrictions, such as it needs to pre-define the network size, its convergence is poor and the structure is not flexible. To overcome these shortcomings, a clustering method based on a growing self-organizing neural network is proposed by the knowledge of self-organizing neural network. This method controls neural's growths and deletions by implementing trigger mechanism of the threshold value without supervision, and through making adjustments of neural weight, it can get clustering results of data objects. The experiment results prove the method's effectiveness and superiority by choosing data objects in two-dimensional space as input samples.

Key words: self-organizing; growing; feature maps; clustering; neural network

0 引 言

聚类是把大量的数据对象按照"物以类聚"的原则划分成若干个类别,使得同一类别内数据对象的相似性尽可能大而不同类别内数据对象的相似性尽可能小[1]。聚类方法有多种,如神经网络法[2]、矢量量化法[3]、C均值法等[4]。Kohonen提出的SOM(Self-Organizing Feature Map)神经网络[5]是一种最主要的神经网络聚类方法,该网络能够模拟大脑神经系统自组织特征映射的功能,通过神经元之间的交互作用和相互竞争,自动地对输入的数据对象进行聚类[6]。但是SOM神经网络用于聚类存在三个主要问题:

(1)网络的大小(即神经元的总数)需要预先设定[7],而在没有先验知识的情况下,很难一次性设定出最佳大小,必须经过不断地尝试修改才能获得较为理想的聚类效果;

(2)网络的学习速率和邻域大小等参数在网络学习过程中会逐渐衰减,因此其初始值的设定对网络的收敛性能有很大的影响;

(3)网络的结构固定,不能灵活调整,对未来数据的学习能力弱,当学习模式较少时,网络的聚类效果取决于输入模式的先后顺序。

针对SOM在聚类分析中的三个缺陷,文中提出了一种生长型自组织神经网络的聚类方法。这种生长型自组织神经网络是以Fritzke提出的成长型单元结构(Growing Cell Structures)作为基础[8],在无监督的情况下通过神经元的生长、删除和权值的调整,使得最终的网络能反映出数据对象的聚类结果。

1 生长型自组织神经网络

1.1 网络拓扑结构

初始网络结构是一个 k 维单形,由 $k+1$ 个结点及两两连接的一些边组成。结点代表神经元,边代表神经元之间的邻接关系。k 的值可以任意选择,例如,当

收稿日期:2010-06-28;**修回日期**:2010-10-04
基金项目:山东省自然科学基金(Y2007G34,Y2006G03);南京邮电大学引进人才基金项目(NY207134)
作者简介:傅 雪(1986-),女,江苏南京人,硕士,研究方向为模式识别与智能系统;张少白,硕士研究生导师,研究方向为人工智能与认知科学、信息获取、处理与识别等。

$k=1$ 时,是一条线段;$k=2$ 时,是一个三角形;$k=3$ 时,是一个四面体。文中只分析较为典型的二维单形网络结构,即 $k=2$ 的情况。图1为网络学习到一定阶段后的拓扑结构。在网络学习的过程中,每个神经元的位置不断调整,网络逐渐生长出新的神经元,并删除"无用"神经元,但是经过任何一次的修改操作,网络结构总是维持一个由一系列三角形组成的二维单形。

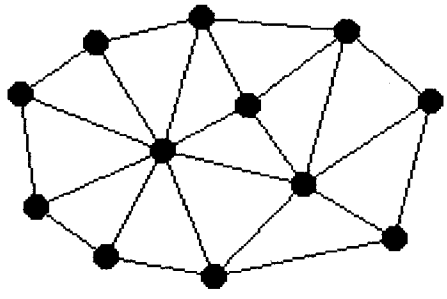

图1　k=2 时的网络拓扑结构

1.2　网络学习规则

网络的学习过程主要包括神经元位置的移动、神经元的生长和神经元的删除[9]。假设输入样本也是二维的,为了使网络能准确学习到输入样本的分布,需要不断地从输入样本中选取一个二维矢量,通常,这种选择是随机的,因此,只有当选取的二维矢量的数目足够多时,网络的学习结果才能达到一定的精确性。

1.2.1　神经元位置的移动

每个神经元都有一个对应的错误变量值,值的大小将决定神经元如何移动位置,也决定着新的神经元的生长位置。每个神经元的错误变量的初始值是预先给定、大小相同的。当选取了输入样本中的一个二维矢量时,网络中将产生一个最佳匹配单元,即与该二维矢量的欧式距离最小的那个神经元。然后使最佳匹配单元与其所有邻接神经元朝着二维矢量所在位置的方向移动,其余神经元位置保持不变。这样,称网络"学习"了一次。期间,最佳匹配单元与其所有邻接神经元的错误变量值也将增加。

1.2.2　神经元的生长

定义一个计数器 Num1 和一个阈值 Threshold1,Num1 用来记录从输入样本中选取的二维矢量数目。当 Num1 = Threshold1 时,网络中将生长出一个新的神经元。此时,Num1 清零,重新开始计数。生长一个新的神经元的规则为:选择当前网络中错误变量值最大的一个神经元 a,并在其所有邻接神经元中选取与其连接权值最大的一个神经元 b,通过分裂 a 与 b 生长出一个新的神经元 i,i 初始位于 a 和 b 的中心点,为了使网络结构维持一个二维单形,相应的边也要作一些增加和删除。同时,新的神经元与其所有邻接神经元的错误变量值将作修改。

1.2.3　神经元的删除

定义一个计数器 Num2、一个获胜计数器 WinNum 和一个阈值 Threshold2,Num2 用来记录当前从输入样本中选取的二维矢量数目,也可以看成是二维矢量编号,WinNum 用来记录每个神经元最近一次成为最佳匹配单元所对应的二维矢量编号。当网络学习了一定次数以后,如果网络中的某个神经元从来没有成为最佳匹配单元或者成为最佳匹配单元的次数极少时,这个神经元就为"无用"神经元,因为它不能代表输入样本的模式,相反,会影响聚类的效果。当 Num2 - WinNum = Threshold2 时,该神经元将从网络中删除,相应的边也随之删除。

1.3　算　法

生长型自组织神经网络的算法记述如下[10]:

Step1 确定网络的维数并初始化网络。这里取 $k=2$,因此初始化的网络是一个三角形。集合 A 包含三个神经元,$A=\{c_1,c_2,c_3\}$。

Step2 从给定的二维输入样本中随机选取一个二维矢量 $\boldsymbol{\xi}$。

Step3 计算 $\boldsymbol{\xi}$ 与每个神经元的欧式距离 $\|\boldsymbol{\xi}-\boldsymbol{\omega}_i\|$($\boldsymbol{\omega}_i$ 为神经元 i 的位置矢量),从中选择最佳匹配单元,即满足 $\|\boldsymbol{\xi}-\boldsymbol{\omega}_s\|=\min\limits_{n\in A}\|\boldsymbol{\xi}-\boldsymbol{\omega}_n\|$。

Step4 修改最佳匹配单元的错误变量:$\Delta E_s=\|\boldsymbol{\xi}-\boldsymbol{\omega}_s\|^2$。

Step5 修改最佳匹配单元及其所有邻接神经元的位置矢量:

$$\Delta\boldsymbol{\omega}_s=\varepsilon_b(\boldsymbol{\xi}-\boldsymbol{\omega}_s)\quad \Delta\boldsymbol{\omega}_i=\varepsilon_n(\boldsymbol{\xi}-\boldsymbol{\omega}_i)\quad (\forall i\in N_s)$$

这里,由于位置矢量一次性修改的幅度不大,所以参数 ε_b 和 ε_n 的取值很小,一般在0.0005至0.2之间,N_s 代表神经元 s 的所有邻接神经元。

Step6 当选取的二维矢量数目 Num1 达到阈值 Threshold1 时,即 Num1 = Threshold1,Num1 清零,并且按照如下方式新增一个神经元:

a)选择网络中错误变量值最大的神经元 q:$q=\arg\max\limits_{c\in A}E_c$。

b)在 q 的所有邻接神经元中,选择距离 q 最远的(即与 q 的连接权值最大)神经元 f。

c)新的神经元 r 的初始位置位于 q 和 f 的中心,即 $\boldsymbol{\omega}_r=(\boldsymbol{\omega}_q+\boldsymbol{\omega}_f)/2$。

d)删除边(q,f),增加边(q,r),(f,r)。找出 q 和 f 的共同邻接神经元,将每个共同的邻接神经元和新的神经元 r 连接。

e)修改 r 的每个邻接神经元的错误变量值:

$\Delta E_i=-\dfrac{\alpha}{|E_N|}E_i$　$(\forall i\in N_r)$,这里,$|N_r|$ 代表神经

元 r 及其所有邻接神经元的总数量。

f)修改 r 的错误变量值,使其为所有邻接神经元的错误变量值的均值:

$$E_r=\frac{1}{N_r}\sum_{i\in N_r}E_i$$

Step7 当 Num2 - WinNum = Threshold2 时,删除这个神经元及所有包含该神经元的边。Num2 为当前从输入样本中选取的二维矢量数目,WinNum 为每个神经元最近一次成为最佳匹配单元所对应的二维矢量编号。

Step8 修改网络中所有神经元的错误变量值:

$$\Delta E_c=-\beta E_c(\forall c\in A)$$

Step9 如果当前网络中的神经元数还未达到规定的最大值 maxNodes,则返回 Step2。

2 聚类分析仿真实验

应用 JAVA 仿真该网络,实现对数据的聚类分析。JAVA 语言是面向对象的,可以将神经元、神经网络模型的属性和运算抽象出来,封装成类。另外,JAVA 语言又具有继承性,可以从最为抽象的神经元、神经网络模型中方便地扩充出各种有自己特点的神经元、神经网络,通过神经元对象之间的消息传递,可以实现神经元之间的侧向交互原理[11]。

如图 2 所示,输入样本向量集中分布在几个形状各异的图形区域[12]。首先初始化网络的竞争层,三个神经元在二维空间内随机分布,初始的网络拓扑结构是一个三角形,设 $\varepsilon_b=0.1$, $\varepsilon_n=0.001$, $\alpha=2.0$, $\beta=0.0005$, max Nodes = 100, Threshold1 = 25, Threshold2 = 500。图 2 所示的是前 11 个新的神经元的生长过程,各图中的白色结点代表新生长的神经元,可以看到,在网络的学习过程中,除了神经元个数逐渐增加外,神经元的位置也在不断调整。输入样本向量越集中的地方,神经元个数越多,因为原来分布在输入样本向量集中区域的神经元成为最佳匹配单元的可能会更大,而且最佳匹配单元和其邻接神经元总是朝着输入样本向量的位置移动。网络在神经元的生长过程中,始终维持一个二维单形的结构。随着网络规模的逐渐扩大,网络中会出现一些"无用"神经元,即游离在输入样本之外的神经元,这些神经元获胜的可能性小,并且会影响到网络的聚类结果,采用上述的阈值控制机制可以有效地删除它们,得到清晰的聚类结果。

图 3 显示了网络经过无监督的自组织、自学习过程,最终形成的聚类结果。图 4 为 SOM 的聚类结果。通过比较两种网络的聚类结果发现,相同的输入样本,

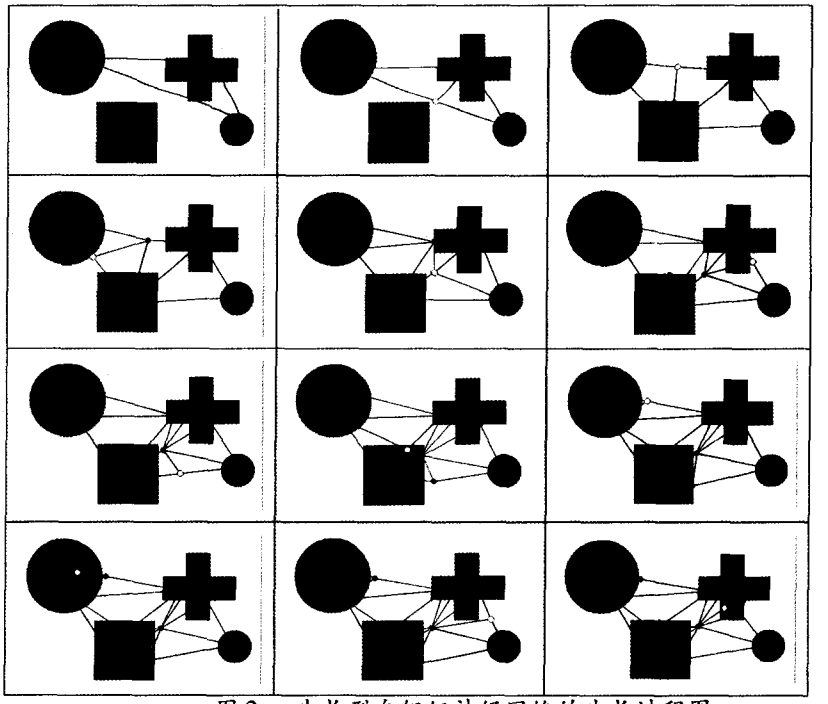

图 2 生长型自组织神经网络的生长过程图

生长型神经网络只要 100 个神经元就能得到较好的聚类结果,而 SOM 网络则需要 225 个神经元,也就是利用 SOM 网络聚类选取的二维输入矢量总数要远远大于生长型神经网络;此外,由于生长型神经网络采用了"删除神经元"机制,因此在聚类效果上要胜于 SOM。

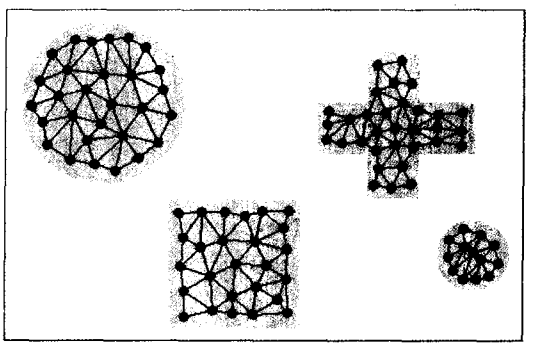

图 3 生长型神经网络的聚类结果

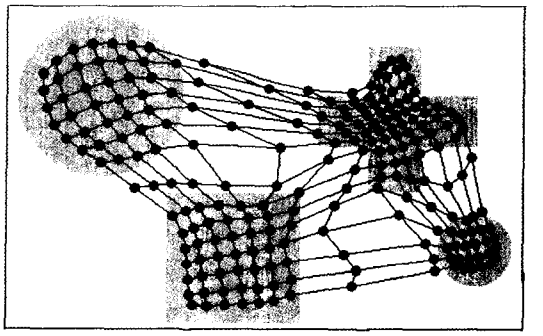

图 4 SOM 的聚类结果

(下转第 71 页)

```
return listResult;
}
}
```

4 结束语

文中介绍了优秀的开源全文检索引擎 Lucene,并在其上根据某军用信息系统的特点和要求进行了二次开发,按照 Lucene 的框架规范,将 Lucene 很好地嵌入到了该系统中,实践证明,该系统达到了预期的目标,全文检索效率高,准确率和查全率都达到了设计的要求。文中提出的解决全文检索系统的方法,也可方便地应用到其他企业的信息系统中,能有利地提高企业对目标文档的管理水平与检索能力。

参考文献:

[1] 林碧英,赵　锐,陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展,2007,17(5):184-185.

[2] 吴　青,夏红霞,赵广辉,等. 基于 Lucene 全文检索引擎的应用与改进[J]. 武汉理工大学学报,2008,30(7):145-146.

[3] 周　宁,谷宏群. 基于 Lucene /XML 的全文检索机制研究[J]. 图书·情报·知识,2005(6):75-76.

[4] Hatcher E, Gospodnetic O. Lucene in action[M]. [s. l.]: Manning Publications Co. ,2005.

[5] 阳　奇,林镇灿,黄　帆,等. 基于 Hibernate 搜索的数据库全文检索系统[J]. 计算机工程,2010,36(4):73-76.

[6] 徐爱春,魏艳华,何震晏. 基于 Lucene 的电子政务全文检索系统的设计与实现[J]. 现代情报,2008,7(7):223-225.

[7] 叶　靓,王智斌,邵谦明. 基于相关反馈的语音检索引擎[J]. 计算机工程,2007,33(17):228-230.

[8] 蒋一峰,王　华,张玉红,等. 基于 Lucene 的语义检索系统的设计和实现[J]. 计算机工程与设计,2008,29(20):5336-5337.

[9] 高　琰,谷士文,谭立球,等. 基于 Lucene 的搜索引擎设计与实现[J]. 微机发展(现更名:计算机技术与发展),2004,14(10):27-30.

[10] 张雪燕,杨晟刚,黄　文,等. 企业级搜索引擎技术在博客网站中的应用[J]. 计算机工程与设计,2008,29(18):4856-4858.

[11] 赵　汀,孟祥武. 基于 Lucene API 的中文全文数据库设计与实现[J]. 计算机工程与设计,2003(20):179-181.

[12] 徐宝文,张卫丰. 搜索引擎与信息获取技术[M]. 北京:清华大学出版社,2003.

[13] 张校乾,金玉玲,侯丽波. 一种基于 Lucene 检索引擎的全文数据库的研究与实现[J]. 现代图书情报技术,2005(2):41-42.

[14] 管建和,甘剑峰. 基于 Lucene 全文检索引擎的应用研究与实现[J]. 计算机工程与设计,2007,28(2):490-491.

(上接第66页)

3 结束语

生长型自组织神经网络的聚类过程是无监督、自学习和自组织的,不需要事先确定聚类中心。相对于 SOM 网络而言,其优越性在于:(1)不需要预定义网络的大小,而是通过神经元的增长来逐渐扩张网络直至达到某种性能或规模大小;(2)该种网络的各项参数都是固定值,不会影响到网络的收敛性,当然参数的取值是否合理会影响到最后的聚类结果,因此需要重复调整参数值的大小来不断观察聚类结果;(3)生长型自组织神经网络结构可以灵活调整,除了可以新增神经元,也可以删除"无用"神经元,对未来数据的学习能力强,这样形成的聚类结果更加清晰。网络利用阈值控制的触发机制有效地实现了神经元的增删,为了进一步提高网络的快速学习能力,如何确定一个相对较优的权值调整系数将是下一步研究的重点。

参考文献:

[1] 陈学进. 数据挖掘中的聚类分析的研究[J]. 计算机技术与发展, 2006, 16(9): 44-49.

[2] 倪步喜, 章丽芙, 姚　敏. 基于 SOFM 网络的聚类分析[J]. 计算机工程与设计,2006, 27(5): 855-878.

[3] 徐　勇, 戴逸松, 陈贺新. 神经网络矢量量化的设计与实现[J]. 长春邮电学院学报,1996, 14(3): 1-4.

[4] 赵　鹏, 耿焕同, 蔡庆生, 等. 一种基于加权复杂网络特征的 K-means 聚类算法[J]. 计算机技术与发展,2007, 17(9): 35-37.

[5] Kohonen T. The Self-organizing Map [J]. IEEE, 1990, 78(9):1464-1480.

[6] 董长虹. Matlab 神经网络与应用[M]. 北京:国际工业出版社, 2007: 185-196.

[7] 傅　彦, 周俊临. 基于自增长型多级自组织映射网络的模式识别[J] 计算机科学, 2004, 31(5): 159-162.

[8] Fritzke B. Growing cell structures - a self-organizing network for unsupervised and supervised learning [J]. Neural Networks, 1994, 7(9): 1441-1460.

[9] 吴　郢, 阎平凡. 结构自适应自组织神经网络的研究[J]. 电子学报, 1999, 27(7): 55-58.

[10] Fritzke B. Growing cell structures- a self-organizing network in k dimensions [M]// Artificial Neural Networks. [s. l.]: [s. n.],1992:1051-1056.

[11] 李　戈,邵峰晶,朱本浩. 基于神经网络聚类的研究[J]. 青岛大学学报, 2001, 16(4): 21-24.

[12] Fritzke B. Unspervised clustering with growing cell structures [C]//Proc. Int. Joint Conf. on Neural Networks. [s. l.]: [s. n.], 1991:531-536.