维普资讯 http://www.cqvip.com

一种基于改进的自组织特征映射网络的文档聚类方法

王升明 李 森
(中国科学院合肥智能机械研究所,合肥 230031)
E-mail:wsm634@163.com

摘 要 采用标准的自组织特征映射(Self—Organizing Feature Map)网络进行文档分类时存在着收敛速度慢、不同的初始条件及学习样本输入顺序影响学习过程和学习结果等缺点。针对这种情况,论文提出了一种改进的自组织特征映射网络文档分类方法,即采用自适应的方法建立网络拓扑结构,利用输入训练样本来确定网络中的连接权值,综合系统能量函数和训练次数作为判断网络的学习结束标准,并且取得了比较好的实验结果。

关键词 自组织特征映射 文档聚类 自适应

文章编号 1002-8331-(2005)03-0167-03 **文献标识码** A **中图分类号** TP301

A Document Clustering Method Based on Improved SOFM Network

Wang Shengming Li Miao
(Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031)

Abstract: Adopting standard SOFM in document clustering, there are some defects including low speed of constringency, different effects on learning and result from different initialization condition and order of input samples. This paper presents a document clustering method which includes some betterments, such as to set up the network structure with self-adaptive method, to determine the initialization weights on input samples, and to determine the end on both system energy and time of learning. At last, it acquires better result in experiment.

Keywords: SOFM, document clustering, self-adaptive

1 引言

信息技术的快速发展使得信息量急剧增长,其表现之一就是各种文档数目的剧增,从而使得有用信息淹没在大量的文档中。如何有效利用文档成了研究的重点,其中发展起来的有信息检索系统、搜索引擎等,这些系统中的一个重要组成部分便是文档的分类。好的文档分类方法能够提高文档组织的合理性,大大减少信息搜索的时间开销,并有利于提高检索精度。如今,传统的人工文档分类方法已经不能应付文档日益的变化,因此必须探索自动分类的方法。国外在文档分类以及相关的信息检索、信息抽取等领域进行了较为深入的研究,取得了不少引人注目的研究成果,同时产生了一些试验性的系统,这些系统的运行结果表明,由 Kohonen 提出的向量空间模型(VSM),是大规模语料库较好的表示模型,1981 年,Kohonen[1]根据神经元有序的排列可以反映出所感觉到的外界刺激的某些物理特征,提出了自组织特征映射(SOFM)神经网络模型。他认为神经网络中邻近的各种神经单元通过侧向交互作用彼此相互竞争,自适应地发展成识别不同类别信息的特殊分类识别器。自组织特征映射神经网络是一种无教师的聚类方法,它能将高维模式映射到平面上,而保持其拓扑结构不变。利用这种网络模型可以实现文档的自动分类,但是传统的自组织特征映射网络存在着收敛速度慢、学习过程和学习结果容易受输入样本的顺序的干扰等缺点。

论文将采用一种改进方法来对传统的 SOFM 算法进行优化。对网络初始连接权值进行一定策略的优化,使得连接权值更符合于特定的模式类,加快了网络学习的速度,同时采用罗立民[2]提出的系统能量函数作为一种新的网络训练结束判断标准。这样解决了采用循环次数作为网络学习唯一的结束判断标准时存在的缺点,如克服了网络学习已经达到了循环次数但网络精度不够,以及精度已经达到要求然而网络依然在进行学习的缺点。

2 自组织特征映射网络及其算法

作为无教师学习的自组织特征映射神经网络能够将高维模式映射到平面上,并且保持其拓扑结构不变。一般的自组织特征映射拓扑结构如图 1(2-D 空间结构)。

网络上层(图中只画出部分)为输出节点(共 m 个),以二维形式排成一个节点矩阵。图中下层节点表示输入节点(共 d 个)。从输入节点到所有的输出节点都有权连接(图中仅画出一部分)。

基金项目:国家 863 高技术研究发展计划项目(编号:2001AA115400)

作者简介:王升明,硕士生,研究方向为人工智能及其应用。李森,研究员,研究方向为专家系统,人工智能及其应用。

维普资讯 http://www.cqvip.com

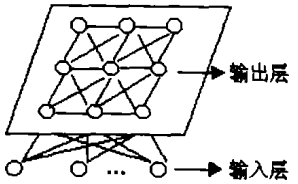

图 1 SOFM 网络拓扑图

自组织的过程实际上就是一种无指导的学习过程，它通过网络自身的训练，自动对输入模式进行聚类。将每个输出单元标记为 $u_i(i=1,2,\cdots,m)$，它接受模式为 $i=(i_1,i_2,\cdots,i_d)^T$ 的输入向量。定义 u_i 的权向量为 $m_i=(w_{i1},w_{i2},\cdots,w_{id})^T$。自组织特征映射的目标就是对于大量的、无标记的训练集，让网络自动去标记输入模式。输出层上相邻的节点能对实际模式分布中相邻的模式类做出特别的反应，能对某一类模式做出特别反应的输出节点代表该模式类。当某类数据模式输入时，对某一输出节点产生最大刺激(获胜节点)，同时对获胜节点周围的一些节点产生较大刺激。在训练过程中，不但要对获胜节点的连接权值进行调整，同时也要对获胜节点的邻域节点的连接权值进行调整，随着训练的进行，这个邻域范围不断缩小。直到最后，只对获胜节点进行细微的连接权值调整。具体的算法如下：

(1)随机初始化连接权值，设定最大训练次数 K，训练计数器 $k=0$。

(2)随机选取输入模式 $i=(i_1,i_2,\cdots,i_d)^T$，计算所有输出单元的欧几里德距离：

$d_j=\|i(k)-m_j(k)\|$

(3)选取获胜节点 u_k，选择标准是：

$\|i(k)-m_u(k)\|=\min\|i(k)-m_j(k)\|\quad(i=1,2,\cdots,m)$

(4)对获胜节点及其邻域节点的连接权值进行调整：

$\Delta m_i=\alpha(k)\times[i(k)-m_i(k)]$

(5)计数器自加，若 $k<K$，转向(2)，否则结束训练。

(6)输出结果。

3 改进的 SOFM 算法

实践表明，对于同一网络和输入样本集，连接权值初始化的好坏明显影响着网络的收敛速度。好的初始权值的设置可以加快网络的收敛，反之亦然。如果输出节点的连接权值完全等同于某一输入节点向量，那么该输出节点就是获胜节点，即该节点可以看作是与输入节点同类的训练样本的最佳样本。罗立民提出可以运用该输入节点向量去初始化输出节点的连接权值。但是，为了使连接权值初始化合理，必须计算输入向量间的欧氏距离，选出 m(m 为输出节点数)个输入向量作为连接权值的标准，这样对于输入样本为 n 的网络，要额外增加计算复杂度为 $O(n^2)$的计算。针对文档聚类的特殊性，为了使网络初始化更加合理，采取了一定的策略对网络连接权值随机初始化进行了优化，并且通过定义系统的能量函数，将系统的能量函数和训练次数结合起来，自适应地调整学习过程。采用系统能量函数作为网络学习结束的一个标准，不仅加快了网络的学习速度，而且加大了网络的聚类精度。系统能量函数定义如下：

$$E(k)=\sum_{i=1}^{n}\sum_{j=1}^{m}\left[w_{ij}(k-1)-w_{ij}(k)\right]^2$$

改进后的算法分两步：

(1)初始化输入样本与输出样本之间的连接权值。由于连接权值的设置合适与否直接影响到网络学习的收敛速度。采用以下策略进行原先算法(随机生成连接权值)的优化：

①设定分类数 m；

②随机将输入样本分成 m 类，计算其聚类重心：

$$C(k)=\frac{1}{n_k}\sum_{X_l\in S(k)}X_l$$

其中，n_k 表示划分子集 $S(k)$的样本数，$k=1,2,\cdots,m$；

③用 $C(k)$初始化网络，并以一定的策略(欧氏距离最大的向量对应于网络上相距最远的输出向量[3])组织网络结构。

(2)网络训练过程

①设定最大训练次数 K，训练计数器 $k=0$；

②随机选取输入模式 $i=(i_1,i_2,\cdots,i_d)^T$，计算所有输出单元的欧氏距离：

$d_j=\|i(k)-m_j(k)\|$

③选取获胜节点 u_k，选择标准是：

$\|i(k)-m_u(k)\|=\min\|i(k)-m_j(k)\|\quad(i=1,2,\cdots,m)$

④对获胜节点及其邻域节点的连接权值进行调整：

$\Delta m_i=\alpha(k)\times[i(k)-m_i(k)]$

⑤调整训练速度：

$$\alpha(k+1)=\alpha(0)\times\left(1-\frac{k}{K}\right)$$

⑥计数器自加，若 $k<K$ 转向②，否则计算网络的能量函数 $E(K)$，如果 $E>\varepsilon$，$K=INT(K\times(1+\zeta))$；转到②，否则到⑦；

⑦输出结果。

算法中 ε 表示系统精度，ζ 表示计数器变化步伐，可根据网络能量的变化进行调整。

由于算法是要适应于文档的分类，但是随着信息技术的发展，文档数量每天都有着变化，所以遇到的另一个问题便是网络结构应该随着新加入文档的变化而变化，并且要把新加入文档的拓扑信息反应到网络中。然后对文档进行类别标注，以便为以后的文档检索服务。所以在完成上面的算法后，在实际的操作中要加入一些后续的操作，使得网络具有自适应性。这里将对网络的进一步学习提出以下改进：

(1)保持第一阶段网络学习的结果，即网络的连接权值不变。设置迭代次数 K。

(2)对网络的调整：

设 k 次迭代输入样本为 $X(k)$，根据最近距离法判断其应获胜的输出接点为 $I_0(k)$。设网络判定 $X(k)$的获胜节点为 $I(k)$，即有：

$$\|X(k)-W_i(k)\|<\|X(k)-W_{i0}(k)\|$$

则可按下列公式进行权矢量的调整以及迭代计算。

①如果 $I(k)=I_0(k)$：

$$w_l(k+1)=\begin{cases}w_l(k)+\alpha(k)[X(k)-w_l(k)], & l=i(k)\\ w_l(k), & \text{其他}\end{cases}$$

②如果 $I(k)\neq I_0(k)$：

$$w_l(k+1)=\begin{cases}w_l(k)+\alpha(k)[X(k)-w_l(k)], & l=i_0(k)\\ w_l(k), & \text{其他}\end{cases}$$

这样，当有新的样本加入时，网络可以自适应地改变连接权值，提高聚类精度。

4 文档的处理

采用词袋(bag-of-word)表示方法来表示文档向量空间。该种表示形式下，忽略文档的具体结构和文档中关键词的顺序，用文档中的关键词组成特征向量。先要去除文档中出现的停用

维普资讯 http://www.cqvip.com

词语(比如“这个”、“一些”,和“的”等等),并且去除训练样本中出现频率不高的词语(或短语)。于是文本可以表示成一个词频向量,它是向量空间中的一个点。于是文档 D 可以表示为向量形式:

$V(D)=(t_1,w_1(D);t_2,w_2(D);\cdots;t_n,w_n(D))$;

其中 t_i 为词条选项,$w_i(D)$ 为 t_i 在文档 D 中的权值。将 $w_i(D)$ 定义成 t_i 的出现频率 $tf_i(D)$ 的函数 $w_i(D)=\psi(tf_i(D))$。ψ 采用 TFIDF 函数:

$$\psi=tf_i(D)\times\log(\frac{N}{n_i})$$

其中,N 为所有的文档数,n_i 为含有词条 t_i 的文档数。这个向量空间往往维数很大,但是真正具有主体区分能力的词条所占比例并不大,多数词条在文本中出现的次数并不多,它们构成了噪声,还有一些出现非常频繁的词条也不具有主题区分能力,这些词条的存在不仅加大了计算复杂度,而且还会淹没主题特征、影响相关文本的聚类。所以应该在具体的文本聚类实现过程中加入了降维处理。降维处理算法[3]如下所示:

设 e_i 表示这样一个向量:除第 I 维的值为 1 外,其余处处为 0,n_{ji} 表示第 I 个项在第 j 篇文档中出现的次数(n_{ij} 可以是其他各种形式的值)。n_j 为第 j 篇文档的向量表示形式,则 n_j 的一种表示方式是:

$$n_j=\sum_k n_{jk}e_k$$

现在,用一个比 e_k 维数低的向量 r_k 代替 e_k,r_k 已被标准化为单位长度。代替后,第 j 篇文档表示为:

$$x_j=\sum_k n_{jk}r_k$$

设矩阵 R 是由向量 r_k 构成的,R 的第 k 列为 r_k。根据上面的两公式,x_j 可以用矩阵乘法来表示:

$x_j=Rn_j$

压缩的实现必须以不改变向量对文档相似性的度量为前提,向量之间的相似性可以用它们的内积来度量:

$x_j^T x_k=n_j^T R^T Rn_k$

$R^T R$ 可以分解为两部分:

$R^T R=I+\varepsilon$

其中:

$\varepsilon_{ij}=r_i^T r_j$

$i=j$ 时,$\varepsilon_{ij}=0$。如果 ε 矩阵的所有分量均为 0,则有 $R^T R=I$。这样,就保持了文档之间相似性。

一种运用统计方法选择 r_k 的方法是 r_k 的各个分量相互独立,来自于期望为 0 的正态分布,将 r_k 的长度标准化为 1。

5 计算机仿真试验结果分析

图 2、3、4 中利用黑点表示文档,其拓扑结构表现了实际存在的文档之间的关系。相似性较高的文档集合在图中就是平面上聚集在一起的黑点。不属于同类的文档之间的欧式距离比较远。利用输出层单元个数为 5×10 的网络进行聚类实验。其结果如图 2、3、4 所示。

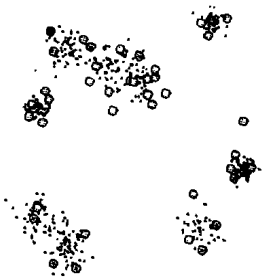

图 2 初始化网络图

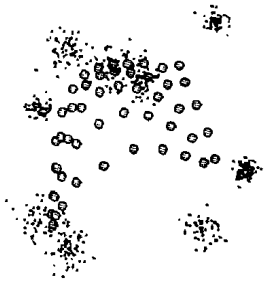

图 3 学习中的网络

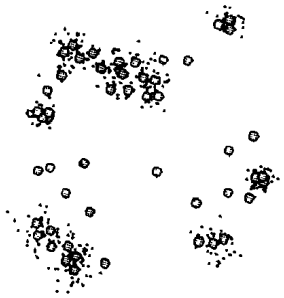

图 4 学习后的网络

网络初始化时由于采取了一定的策略,所以输出点的分布接近于输入聚类文档的分布,但是这并不能真正反映输入文档间的关系,其中有些输出点相近,并且可能出现重叠、聚类空间分布不均匀。算法中初始化的处理加快了网络的收敛速度。图 4 为学习后网络趋于稳定的节点分布情况。这里可以看出输出节点的分布较好地反映了输入节点的拓扑结构,并且分类分布比均匀,达到了预期效果,在每个输出节点的较小邻域内的输入节点可以分为一类。同时实验结果表明经过改进后的算法在时间和精度上都有一定的改观。

6 结束语

论文着重描述了一种改进的 SOFM 算法,并且在文档分类中得以成功应用。改进后的算法加快了自组织特征映射网络的收敛速度,提高了网络的分类精度。对于建立更高效的下一代智能信息检索系统以及智能搜索系统具有一定的借鉴意义。

(收稿日期:2004 年 8 月)

参考文献

1.Robert J Schalkoff.Artificial Neural NetWorkrs[M].MIT Press and The McGraw-Hill Companies,Inc,1997

2.罗立民,王允诚.自组织特征映射网络的改进及在储层预测中的应用[J].石油地球物理勘探,1997;(32):237~245

3.潘文锋.自组织映射(SOM)神经网络及其应用.http://www.software.ict.ac.c/y Seminar/lectures,2003-09

(上接 158 页)

2.K Czajkowski,I Foster,N Karonis et al.A resource management architecture for metacomputing systems[C].In:The 4th Workshop on Job Scheduling Strategies for Parallel Processing,1998:62~82

3.I Foster,A Roy,V Sander.A Quality of Service Architecture that Combines Resource Reservation and Application Adaptation[C].In:International Workshop on Quality of Service,2000

4.I Foster,A Roy,V Sander et al.End-to-End Quality of Service for High-End Applications[R].Technical report,Argonne National Laboratory,Argonne,http://www.mcs.anl.gov/qos/qos papers.htm,1999

5.I Foster,C Kesselman,J Nick et al.The physiology of the grid:An open grid services architecture for distributed systems integration[R].Technical report,Globus Project,http://www.globus.org/research/papers/ogsa.pdf,2002