◎数据库与信息处理◎

基于构造型神经网络和商空间粒度的聚类方法

徐 银[1,2],周文江[1,2],王伦文[2]
XU Yin[1,2],ZHOU Wen-jiang[1,2],WANG Lun-wen[2]

1.解放军电子工程学院 研二队,合肥 230037
2.解放军电子工程学院 309室,合肥 230037
1.The Second Team of Postgraduate Department at Electronic Engineering Institute,Hefei 230037,China
2.309 Research Division of Electronic Engineering Institute,Hefei 230037,China
E-mail:xuyin05657856700@163.com

XU Yin,ZHOU Wen-jiang,WANG Lun-wen.Clustering method based on Constructive Neural Networks and quotient space granularity.Computer Engineering and Applications,2007,43(29):165-167.

Abstract: In this paper,Constructive Neural Networks (i.e.CNN) are used to cluster large-scale patterns,and the optimum granularity is chosen by quotient space granularity analysis method.This method not only makes good use of the characteristic of CNN in reducing the computing complexity,but also takes the advantage of quotient space theory in choosing the optimum granularity. The results of the experiments of clustering large-scale and complicated data show the validity of this method.
Key words: clustering;granularity;constructive neural networks;quotient space

摘 要:采用构造型神经网络对大规模模式进行聚类,其中利用商空间粒度分析法选择最优粒度聚类。该方法既发挥了构造型神经网络计算复杂度低的优点,又利用了商空间理论选取最优粒度聚类。对大规模复杂数据聚类实验结果表明该方法是实效的。
关键词:聚类;粒度;构造型神经网络;商空间
文章编号:1002-8331(2007)29-0165-03 **文献标识码**:A **中图分类号**:TP18

1 引言

在机器学习中,聚类是一个重要的研究课题,聚类是将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。由聚类所生成的簇是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象相异。目前在文献中存在大量的聚类算法,如划分聚类法、密度聚类法、层次聚类法、网格聚类法、模型聚类法、模糊聚类法,以及由这些方法适当组合而成的新的方法, 如:CURE法是一种自底向上的层次聚类算法,首先将输入的每个点为一个聚类,然后合并相似的聚类,直到聚类个数为k,CURE中取固定数目的点来表示一个聚类, 从而提高算法挖掘任意形状的聚类能力;DBSCAN与CURE法相似,不过是对网络上进行聚类,算法可挖掘任意形状的聚类;CURD是对CURE方法的一种改进, 将CURE中"每一类取固定数目的参考点"改为"每一类取(不固定数目)参考点";CLIQUE是一种基于网络和密度的聚类算法。

然而,随着信息技术的发展,Inter网、海量数据库等出现,现有的聚类方法越来越不适应大规模数据的聚类。根据大规模数据的特点, 将构造型神经网络与商空间粒度分析法相结合,提出构造型神经网络与商空间粒度相结合的聚类算法,对大规模数据进行聚类。

文章第二部分简介构造型神经网络与商空间粒度分析理论;第三部分根据聚类的特点,提出构造型神经网络与商空间粒度相结合的聚类算法;第四部分说明具体的聚类算法实现步骤;第五部分将该方法对无线电监测数据进行聚类,验证了算法的实效性。

2 构造型神经网络与商空间粒度分析理论简介

2.1 构造型神经网络简介

构造型神经网络是张铃教授和张钹院士,在M—P神经元几何意义的基础上[1]提出来的。该算法的主要思想是将n维向量投影到$n+1$维球面上,把神经网络的设计问题转化成某种求领域覆盖的问题。具体地说,就是将n维空间中的有界集合D作变换$T:D\to S^n$,映射到S^n是$n+1$维空间中的超球面上。其中,变换关系为:$T(x)=(x,\sqrt{(d^2-|x|^2)})$,$d\geqslant\max\{|x|,x\in D\}$。通过变换,将$D$一一映射到半径为$d$的超球面上,进而在超球面上对数据进行分析和处理。从而构造一个网络,就等价于求出一组领域,对给定样本集的点,能够用领域覆盖将它们分隔开来。这

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60475017);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318108)。
作者简介:徐银(1983-),男,汉族,硕士研究生,研究方向为智能信息处理理论及应用;周文江(1981-),男,汉族,硕士研究生,研究方向为智能信息处理理论及应用;王伦文(1966-),男,博士,副教授,硕士生导师,研究方向为计算智能、机器学习等。

样,就将神经网络的最优设计问题转化成某种求最优覆盖的问题。详细内容可参见文献[1],文献[2]。

构造型神经网络由于几何意义明显、直观、计算量小,使大规模数据的处理问题变得很直观,得到了广泛的关注,对这方面的研究也很多。文献[3]详细介绍了构造型神经网络理论并引进模糊算法对其进行了改善,由此提出了构造型模糊神经网络。文献[4]在文献[3]的基础上,结合粗糙集理论对大规模模式进行分类。文献[5]提出一种将神经网络覆盖算法与模糊集思想相结合的方法,构造了模糊分类器。文献[6]在球邻域模型的基础上提出一种可用于大规模模式识别问题的神经网络训练算法,并应用于手写体汉字的识别。目前,构造型神经网络在大规模分类中已经得到广泛的认可,但是在聚类,特别是大规模数据聚类应用还不够。

2.2 商空间粒度分析理论简介

在讨论一个问题时,往往从不相同的粒度上加以观察和分析。张铃等人由此提出了问题求解粒度的概念[7],具体可描述如下。

将所研究的问题用一个三元组(X,F,Γ)[7]来加以描述,其中X表示问题的论域,也就是要考虑的基本元素的集合。并设F是属性函数,定义为$F:X\to Y$,Y表述基本元素的属性集合。Γ表示论域的结构,定义为论域中各个基本元素之间的关系。

从一个较"粗"的角度看问题,实际上是对X进行简化,把性质相近的元素看成是等价的,把它们归入一类,整体作为一个新元素,这样就形成一个粒度较大的论域$[X]$,从而把原问题(X,F,Γ)转化成新层次上的问题$([X],[Y],[\Gamma])$。粒度和等价关系有着非常密切的联系。

实际上,这种简化过程和拓扑商集的概念完全相同。

定义 1 给定论域X,设R是X上的一个等价关系,令$[x]=\{y|yRx,y\in X\}$,$[X]_R=\{[x]|x\in X\}$,称$[X]_R$为X对应于R的商空间,记为$[X]_R$(或$[X]$),其中xRy表示x与y等价。

定义 2 设R表示由X上一切等价关系的全体。可以如此定义等价关系,也就是粒度的"粗"和"细"。设$R_1,R_2\in R$,若$x,y\in X$,$xR_2y\Rightarrow xR_1y$,则称R_2比R_1细,记为$R_1<R_2$。反之,则称R_2比R_1粗。

定理 1 R在上面定义的"<"关系下形成一完备的半序格[8]。

于是可以得到关系式:$R_0<R_1,\cdots,<R_{n-1}<R_n$。式中$R_n$是最细的等价关系,$R_0$是最粗的。显然,当$R_0<R_1$时,则$[X]_{R0}$是$[X]_{R1}$的商集,即$[X]_{R0}$对应的粒度比$[X]_{R1}$的粗。其中,$[X]_R$表示$X$关于等价关系$R$的商空间。

定理 2 设连通空间X,对于等价关系R,有商集Y,即:$f:X\to Y$是从连通空间到拓扑空间Y的一个连续映射,则$f(X)$是Y的一个连通子集。反之,如果$f(X)$是Y的一个非空隔离子集,X则不连通。

定理 2 表明,若一个问题在原论域X上有解(是连通的),在粗粒度论域$[X]$上也有解。反之,若粗粒度论域无解,则原问题必无解(不连通)。由于粗粒度的世界通常比较简单,求解问题可以先从粗粒度入手,往往得心应手,粗粒度的分析结果又可以指导我们进行细粒度分析,避免求解过程中走弯路。这也是采用商空间粒度进行分析的原因。

3 构造型神经网络与商空间粒度相结合的聚类算法

决定聚类结果的主要因素有两个:一是相似度函数,二是相似度阈值。相似度函数与相似度测量标准有关,常见的标准有点积、距离、相似性比、布尔"与"运算、规范化的相关系数、模糊等价关系等。不同的相似性度量标准对应不同的聚类算法,本文利用构造型神经网络聚类采用内积作为相似性准则。详细内容参见文献[9]。

相似度函数确定后,聚类的结果由相似度阈值决定。被划分类别的大小和多少直接与阈值有关。聚类的结果一般都使用聚类谱系图来表示[8]。比如对于图 1 中的 4 个样本点,相应的聚类谱系图为图 2。从聚类谱系图中可看出,如果选取的聚类阈值R足够大的话($R>R_3$),那么所有的样本点都被归为一类;如果$R_2<R<R_3$,那么所有的样本点被分为两类,样本点{3,4}归为一类,剩余的样本点归为另一类。随着R的减小,类数越来越多,直到所有的样本点自成一类。

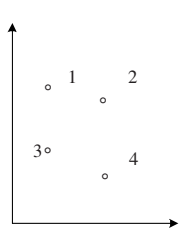

图 1 聚类的样本点

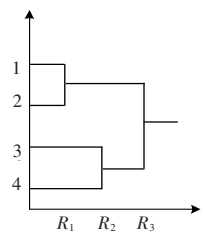

图 2 图 1的谱系图

由此可见,可以用粒度描述聚类的粗细。对于给定的相似度函数,取不同的阈值,必然得到一聚类,这些聚类一般是不同的。当采用较大的阈值时,展现在我们面前的是样本点集比较"粗"的轮廓,一些细枝末节被忽略掉了;而采用较小的阈值时,就能够比较精细地刻画样本点之间一些细微差别。比如图 2 中,当阈值$R>R_3$时,所有样本被聚一类,称粗粒度聚类,而$R<R_1$时,所有样本各成一类,称细粒度聚类。

因此,聚类和粒度之间存在密切的联系,实际上,聚类分析是以"最优"相似度函数为基础,在所有可能的粒度中,寻找出一个"最优"粒度。为了寻求合适的粒度,下面考察两个有益的等价划分。

定义 3 设R_1和R_2是论域 X 上的两个等价关系,如果R也是X上的一个等价关系,并且同时满足下面两个条件,那么称R为R_1和R_2之积,记为$R=R_1\otimes R_2$。

$R_1<R$且$R_2<R$;若还有R',使得$R_1<R'$,$R_2<R'$,且$R<R'$。

定义 4 设R_1和R_2是论域X上的两个等价关系,如果R也是X上的一个等价关系,并且同时满足下面两个条件,那么称R为R_1和R_2之和,记为$R=R_1\oplus R_2$。

$R<R_1$且$R<R_2$;若还有R',使得$R'<R_1$,$R'<R_2$,且$R'<R$。

根据上述两个定义,$R_1\otimes R_2$是能细分R_1和R_2最粗的划分,$R_1\oplus R_2$是能细分R_1和R_2最细的划分。即$R_1\otimes R_2$是划分R_1和R_2的最粗的上界,$R_1\oplus R_2$是划分R_1和R_2的最细的下界。

对于一个具体问题聚类分析时,可参考图 3 所示方法。首先,根据问题需要预置一个等价关系R_0划分问题对应的集合(相应的粒度为Δ_0),得到商空间S_0,在S_0上分析问题,得出初步结论A_0。若根据某判据发现粒度偏粗,这时可(以A_0为指导)取一粒度R_0',令$R_1=R_0\otimes R_0'$,在R_1上再进行分析,得出结论A_1。如果还是粗的话,(以A_1为指导)再取一粒度R_1',令$R_2=R_1\otimes R_1'$,在R_2上再进行分析。以上过程可以重复进行,每重复一次,粒度将细化一次。

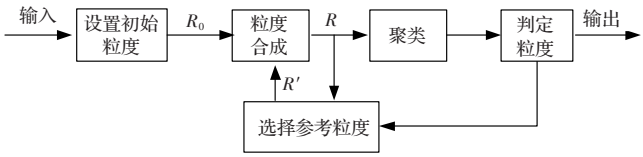

图 3　商空间粒度聚类原理框图

同理,用等价关系 R_0 进行划分,如果粒度偏细,取一粒度 R_0',令 $R_1=R_0\oplus R_0'$,再进行分析。如果还是细的话,再取 R_1',令 $R_2=R_1\oplus R_1'$,再进行分析。以上过程也可以重复进行,每重复一次,粒度将加粗一次。

粗化和细化的过程,根据需要还可混合使用。最终,总是可以找到合适的聚类粒度。

4　算法实现步骤

算法 1　利用覆盖算法聚类

步骤 1　求所有未学习样本的重心，并以离该重心最近的样本作为覆盖的圆心;

步骤 2　求出剩下未聚类的样本与圆心的距离，并求出所有距离的平均距离;

步骤 3　以该平均距离作为初始聚类粒度，求出球形覆盖及球形覆盖的重心;

步骤 4　重新确定圆心,重复步骤 2、3,获得新的覆盖,根据聚类结果判别聚类粒度是否最优,如果最优转至步骤 5,否则利用算法 2 选择粒度,直到覆盖的样本数不再变化为止;

步骤 5　找离当前覆盖的圆心最远的点作为下一步覆盖的圆心;

步骤 6　重复步骤 2–步骤 5,直到所有的样本全部覆盖结束;

步骤 7　若有覆盖包含点较少,需合并覆盖,则采用最短距离法,转至步骤 8;

步骤 8　计算出两覆盖的圆心的距离，将离得最近的两个覆盖合并为一个新的覆盖;

步骤 9　更新其他覆盖与新覆盖的最短距离;

步骤 10　根据实际情况,重复步骤 8,9,确定最后的聚类数。

算法 2　利用商空间粒度进行粒度分析

步骤 1　在初始粒度 Δ_0 对应商集 S_0 上得出初步结论 A_0,根据此结论、粒度 Δ_0,判定聚类粗细,若偏粗转至步骤 2,否则转至步骤 3;

步骤 2　以当前结论为指导，取一相关（偏细）等价关系 R_0',令 $R_1=R_0\otimes R_0'$,在 R_1 上再进行分析,得出聚类粒度 Δ_1 和结论 A_1。若粒度合适,则转至步骤 4,若粒度仍较粗,按照同样方法重复步骤 2;

步骤 3　以当前结论为指导，取一相关（偏粗）等价关系 R_0',令 $R_1=R_0\oplus R_0'$,在 R_1 上再进行分析,得出聚类粒度 Δ_1 和结论 A_1。若粒度合适,则转至步骤 4,若粒度仍较细,按照同样方法重复步骤 3;

步骤 4　分析结束。

5　实验和结论

这里对无线电监测数据进行聚类,通过计算机遥控某短波接收机,使其在 15 MHz 至 16 MHz 频段内,按照一步进,从低到高循环搜索。搜索到任一频点时从接收机的中频输出端采集信号,并经快速傅立叶变换后转化为频域数据。该波段电磁环境非常复杂,截获频域数据数据量大,噪声多,这里对接收的数据进行聚类分析。

图 4 所示为连续 3 天的频域数据占度图。其中横轴代表时间,以天为单位,纵轴代表频率,以 MHz 为单位,图中点的实虚代表该点对应的时间和频率上信号的有无。图 5 为经过覆盖算法聚类后的时频图,图 6 为构造型神经网络和商空间相结合算法(改进算法)聚类的时频图。

通过对实验结果数据分析(如图 4、图 5、图 6 所示)比较,可以发现,监测数据经过覆盖算法聚类后能够较为有效地消除其中的噪声,聚类效果较好。而结合商空间粒度分析的方法后,改进的算法能够更好地显示信号的本来面貌,聚类效果进一步提高。

为了更好地说明算法的有效性，选取划分法中的 k 均值法,层次法中 CURE 法,以及常规覆盖算法做了比较实验,结果如表 1 所示。

表 1　不同聚类方法实验结果

聚类方法	聚类时间/s	总频点数	聚类错误频点数	聚类错误率/%
k–均值法	2.8	333	18	5.41
CURE 法	2.5	333	13	3.90
覆盖算法	2.3	333	10	3.00
改进算法	2.0	333	5	1.50

从表 1 可以看出,采用构造型神经网络和商空间粒度相结合的方法,不仅在聚类时间上比其他方法少,而且聚类的错误率也大大地下降。

以上以内积作为相似度,采用构造型神经网络对无线电监测数据进行聚类,在聚类过程中利用商空间粒度分析法选择最优粒度。从实验结果可以发现，该方法与常见的 k–均值法和 CURE 法及常规覆盖算法相比,不仅聚类速度得到较大幅度提

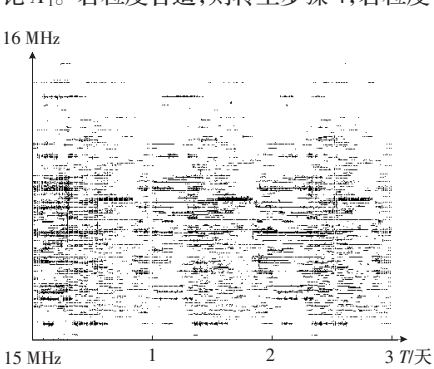

图 4　原始时频图

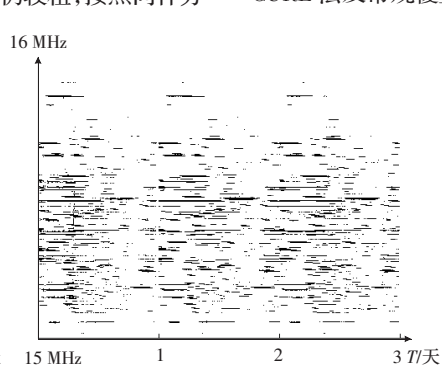

图 5　经覆盖算法聚类后时频图

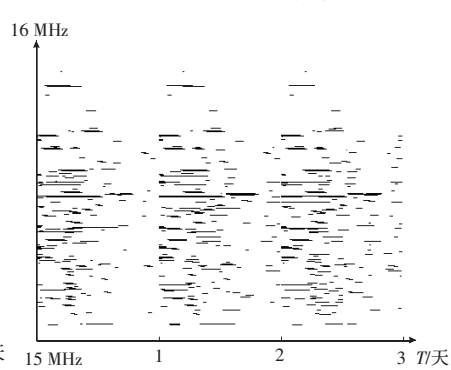

图 6　经改进算法后时频图

（下转 185 页）