文章编号:1671-8860(2012)02-0215-05　　**文献标志码**:A

利用二型模糊聚类进行全球海表温度数据挖掘

孔令桥[1]　秦　昆[1]　龙腾飞[2]

(1　武汉大学遥感信息工程学院,武汉市珞喻路 129 号,430079)
(2　中国科学院对地观测与数字地球科学中心,北京市海淀区邓庄南路 9 号,100094)

摘　要:基于二型模糊集的 C 均值聚类方法对全球时序海表温度数据进行了聚类分析,得到全球海表温度异常的典型聚类模式,并从聚类中心挖掘出潜在的海洋气候指数。

关键词:模糊聚类;海表温度;数据挖掘;海洋气候指数;气候变化
中图法分类号:P208

海表温度(sea surface temperature,SST)是地球气候系统的重要指标,对全球海表温度聚类可以发现全球海表温度异常的聚类模式,有利于研究海陆气候变化的关联关系。利用空间数据挖掘和知识发现可以从数据库中自动或半自动地挖掘事先未知却潜在有用的空间模式[1]。国内外学者对海表温度进行聚类分析的研究还不多[2-6]。Vipin Kumar 教授领导的研究组使用 K 均值聚类得到全球海表温度聚类模式[2],并使用共享最近邻(shared near neighbors,SNN)聚类方法发现潜在的海洋气候指数[7]。而时序海表温度聚类中存在诸多不确定性,需要使用考虑不确定性的聚类分析方法。在硬聚类的迭代过程中,每个数据对象对聚类中心的计算起到了同等的作用,由此得到的聚类中心很难反映类别的典型特征。因此,使用硬聚类方法进行海表温度聚类分析虽然可以发现全球海表温度异常在空间上的大致分布,却难以得到更具典型性的聚类模式。模糊聚类能够得到样本属于各个类别的不确定性程度[8],表达类别归属的模糊性,将其应用于海表温度聚类中能得到全球海表温度的典型聚类模式。相对于一型模糊集,二型模糊集引入次隶属度衡量隶属度的不确定性,能够处理更高阶的模糊性,意味着具有更强的处理不确定性问题的能力[9]。Hwang 和 Rhee 使用两个模糊加权因子 m_1 和 m_2 构造区间二型模糊集,在模糊 C 均值(FCM)算法的基础上将隶属度扩展为一个区间,提出基于区间二型模糊集的 FCM 聚类算法(IT2FCM)[10],成功处理了 FCM 中模糊加权因子 m 的不确定性。IT2FCM 与 FCM 的聚类过程基本类似,不同之处在于迭代过程中将隶属度从一个确定的值变为一个区间,并使用计算区间二型模糊集质心的方法计算聚类中心,其类别判断的过程则是一个降型和解模糊的过程。在数据结构较复杂、类别的形态和密度未知时,使用基于二型模糊集的 C 均值聚类方法比传统的 FCM 聚类方法更加有效。因此,本文采用基于区间二型模糊集的 C 均值(IT2FCM)聚类方法对时序海表温度数据进行数据挖掘。

1　全球海表温度聚类分析

1.1　时序海表温度的去季节性处理

一般来说,地球科学家对一些非季节性变化规律更加感兴趣[2],而时序数据中存在起主导作用的季节性变化特征掩盖了其他的变化规律。另外,时序数据具有较强的时间自相关性,将对聚类分析产生不利影响。去季节性处理可以从时序数据中去除季节性特征,也可以从一定程度上减少时间自相关[3]。如图 1(a)所示为原始的时间序列,可以看到季节性变化非常明显。

本文使用零均值规范化方法对时序数据进行去季节性处理。该方法对于各月的数据计算该月的多年数据的均值和标准差,然后用原始数据减

收稿日期:2011-12-15。
项目来源:国家 973 计划资助项目(2006CB701305)。

去该月的均值并除以标准差,得到规范化处理后的数据。图1(b)为使用零均值规范化方法进行去季节性处理后的海表温度时间序列,可见去季节性处理消除了季节性因素和数据中的时间自相关。

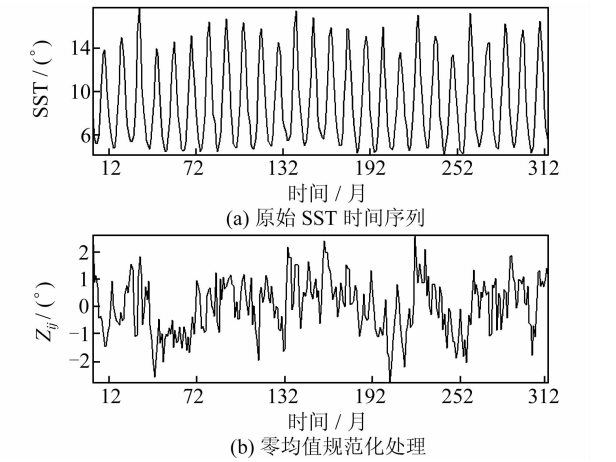

图1 零均值规范化处理

Fig. 1 Monthly Z Score

1.2 时序海表温度聚类的不确定性处理

在全球海表温度聚类中存在许多不确定性。首先,数据中存在大量噪声,如高纬度地区的海洋长年处于极低温状态,海表温度基本保持不变,因此海表温度异常为零,这些地区的时序海表温度数据无法反映海表温度异常,此外还存在其他偶然因素造成的噪声数据。其次,有些格网点属于孤立点,并不属于某一种群聚模式,使用硬聚类方法很难将其分离。因此,在全球时序海表温度数据中存在大量无法反映典型聚类特征的数据。针对时序海表温度聚类中的不确定性问题,采用模糊聚类的方法进行聚类分析。为每个格网点上的时序海表温度赋予隶属度,用来衡量其在归属类别中的典型性,那些具有较高隶属度的时序海表温度具有较高的典型性,更有价值。因此,使用模糊聚类可以得到全球海表温度的典型聚类模式。

1.3 时序海表温度的聚类分析

采用1981~2007年的月平均全球最优插值海表温度NetCDF数据,该数据为经纬度1°×1°全球格网点数据。另外,本文还收集了PDO、ANOM1+2、ANOM3、ANOM3.4、ANOM4等几种基于海表温度异常的海洋气候指数进行辅助分析。

在聚类分析中,常用欧氏距离法衡量样本之间的相似性。然而,时序海表温度聚类分析的对象为经过去季节性处理后的时间序列,其反映海表温度异常的变化,因此这些时间序列之间的相似性应主要体现变化特征的相关,对时序海表温度的聚类就是寻找具有相关变化特征的时间序列的过程。因此,使用相关系数衡量时序数据间的相似性,这种相似性度量方法考虑了时序数据的全局变化特征[11]。式(1)为时序海表温度的皮尔森相关系数的计算公式[2]:

$$r=\sum_i (x_i-\bar{x})(y_i-\bar{y})/\sqrt{\sum_i (x_i-\bar{x})^2\sum_i (y_i-\bar{y})^2} \tag{1}$$

其中,$\boldsymbol{X}=(x_1,x_2,\cdots,x_{312})$,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_{312})$,分别为两个不同格网点上的去季节性后的海表温度时间序列;x_i 和 y_i 是时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 对应时间 i 的海表温度值;$\bar{x}$ 和 $\bar{y}$ 分别为两个时间序列的均值;r 是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关系数,$-1\leqslant r\leqslant 1$。如果聚类对象之间的相关系数较大,则表明这些区域海表温度的变化规律之间具有较高的相似性,通常将这些对象聚成一类。

采用零均值规范化方法对海表温度时序数据进行去季节性处理,使用时序海表温度之间的相关系数作为聚类分析中的相似性度量方法,在此基础上使用基于区间二型模糊集的 C 均值聚类方法将全球时序海表温度聚为50类,如插页Ⅲ彩图1所示。其中,水平轴和垂直轴分别表示经度和纬度,右侧的颜色条显示的不同颜色表示不同的类,属于同一类格网点上的海表温度的变化的相关性较强。

聚类分析的结果可以反映全球海表温度异常的群聚模式。模糊聚类使用隶属度衡量类属的典型性,在时序海表温度的模糊聚类中,隶属度越大,该时间序列更能典型代表其所属类别的特征。因此,在插页Ⅲ彩图1所示的聚类结果的基础上,保留隶属度较大的格网点进行显示,由此得到的海表温度群聚模式能体现全球海表温度异常的典型分布。插页Ⅲ彩图2所示为设置隶属度阈值为0.5的聚类结果,图中用不同颜色显示不同类别,相同颜色表示的块状区域表示一种聚类模式。通过模糊聚类,在世界上许多区域形成了海表温度异常的聚类模式,每种聚类模式代表一种典型的海表温度异常现象。通过模糊聚类得到的聚类中心能够总体反映聚类模式所覆盖区域的海表温度异常的时间序列。

使用插页Ⅲ彩图2所示的典型聚类模式,将聚类中心与海洋气候指数进行相关分析。插页Ⅲ彩图3显示了几个与厄尔尼诺指数(ANOM1+2、ANOM3、ANOM3.4、ANOM4)相关性较大的类别,插页Ⅲ彩图3(c)中用颜色显示类31,其聚类中心与ANOM1+2的相关系数达到0.933 8,与ANOM3的相关系数达到0.881 9;彩图3(e)中

用颜色显示类 17,其聚类中心与 ANOM3 的相关系数达到 0.840 5,与 ANOM4 的相关系数达到 0.926 5,与 ANOM3.4 的相关系数达到 0.945 1。

插页Ⅲ彩图 3 中显示的相关分析结果能够验证海表温度模糊聚类分析的有效性。众所周知,厄尔尼诺常常会使北美地区当年出现暖冬,南美沿海持续多雨;英国的科学家曾指出,拉尼娜现象将使北美洲的西部地区、南美洲及非洲东部地区面临干旱威胁;厄尔尼诺现象往往造成南美西岸持续大雨、东南非洲大范围干旱,并常给北美西岸地区造成频繁的强风暴活动[12]。插页Ⅲ彩图 3(a)、3(b)所显示的类覆盖了北美西岸,彩图3(c)显示的类覆盖了南美洲的西海岸,彩图 3(d)则显示了非洲东部海岸,这些相关分析的结果验证了已知的与厄尔尼诺现象有关的地区。同时,插页Ⅲ彩图 3(c)和 3(e)所示的类别与厄尔尼诺指数的相关系数达到了相当高的水平,而这两个类别所覆盖的区域与定义厄尔尼诺指数的地理位置基本一致,可以认为是对厄尔尼诺指数的重现。插页Ⅲ彩图 3(f)、3(g)、3(h)分别显示了通过模糊聚类得到的类 31 与 ANOM1+2 指数、类 17 分别与 ANOM4 指数和 ANOM3.4 指数的时间序列曲线图,可以看到,它们的相似度很高,其变化趋势基本一致。

相比 K 均值聚类,本文所采用的基于二型模糊集的 C 均值聚类方法能够得到更具典型性的 SST 聚类中心。分别使用模糊聚类和 K 均值聚类得到 SST 聚类中心,与各厄尔尼诺指数进行相关分析,表 1 给出了两种方法得到的两组聚类中心分别能达到的最高相关系数,可以看到,使用模糊聚类得到的聚类中心能够与厄尔尼诺气候指数达到更好的相关性,这在插页Ⅲ彩图 3 中的时间序列曲线图中也得到了体现。

表 1　聚类中心与气候指数的最大相关性比较

Tab. 1　Cormparison of the Maximum Correlation of Clustering Centers and Climate Indices

	PDO	ANOM1+2	ANOM3	ANOM3.4	ANOM4
K 均值聚类	−0.651 6	0.903 7	0.865 1	0.881 3	0.826 0
二型模糊 C 均值聚类	−0.777 9	0.933 8	0.881 9	0.945 1	0.926 5

综上所述,相比传统的硬聚类方法,使用基于二型模糊集的 C 均值聚类方法对全球时序海表温度进行聚类分析更加有效,得到的聚类中心可以反映全球海表温度异常的典型分布,利用这种典型聚类模式研究全球气候变化更有意义。

2　挖掘海洋气候指数

Vipin Kumar 教授领导的研究组使用 SNN 聚类方法发现了一些潜在的海洋气候指数[3]。SNN 聚类方法可以自动获取聚类数目,并且得到地理上连续的聚类。然而,这种方法不能直接处理时序海表温度聚类中的不确定性,往往得到数目繁多的聚类模式,需要进行大量的筛选。本文使用模糊聚类方法可以直接得到具有典型性的海表温度聚类模式,从得到的聚类中心中挖掘出新的海洋气候指数。

在全球较大范围内,如果 SST 聚类中心与陆地气温的相关性更大,则此 SST 聚类中心可以作为已知海洋气候指数的变体,或者可以视为潜在的候选海洋气候指数。因此使用全球时序陆地气温数据作为辅助,计算各海表温度聚类中心与全球陆地气温的相关系数,从而找出对陆地区域影响较广的 SST 聚类中心,在此基础上,比较 SST 聚类中心与海洋气候指数对陆地气温的影响大小,从而挖掘出潜在的海洋气候指数。实验框架如图 2 所示。

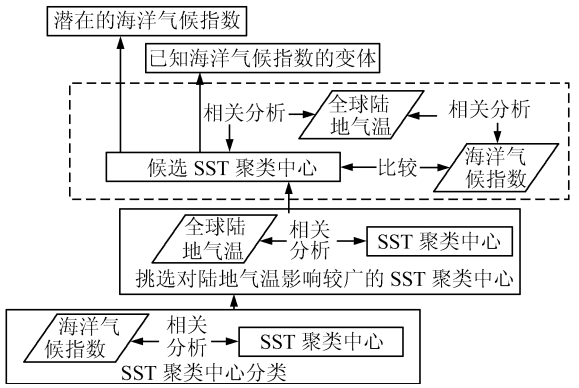

图 2　发现海洋气候指数实验框架

Fig. 2　Frame of Ocean Climate Indices Discovery

2.1　挑选候选聚类中心

海洋气候指数对陆地气候具有指示作用,与陆地气温相关性较好,插页Ⅲ彩图 4 显示了厄尔尼诺指数 ANOM4 与全球陆地气温的相关性,每个陆地格网点对应一个相关系数值,右侧的颜色条指示相关系数的大小,横纵坐标分别为经度和纬度。可以看到,ANOM4 指数与地球上某些区域陆地气温的相关性较好,特别是受厄尔尼诺现象影响较大的区域,如南美洲西南部、非洲东北部、东南亚和澳大利亚北部等。

为了探知海表温度聚类中心对陆地气温的影响范围,计算海表温度聚类中心与陆地气温的相

关系数。如插页Ⅲ彩图4所示,与海洋气候指数的相关系数大于0.2的地区的相关性相对更显著,因此统计与每个SST聚类中心的相关性大于0.2的陆地格网点的个数,得到的直方图如图3所示,横坐标为SST聚类中心的标号,纵坐标为与SST聚类中心的相关性大于0.2的陆地格网点个数。为了评判SST聚类中心的影响范围的大小,将已知海洋气候指数作为参照,故将海洋气候指数也列入直方图,横坐标中的51、52、53、54分别代表PDO指数、ANOM1+2指数、ANOM3指数、ANOM4指数。从图3可以看出,海洋气候指数对应的陆地格网点数均大于300。对SST聚类中心来说,与陆地气温具有较大相关性的陆地区域越大,该聚类中心越有可能成为潜在的气候指数。因此,挑选出图中对应陆地格网点数大于300的聚类中心,认为这些聚类中心对陆地气候的影响较广,作为后续分析的候选SST聚类中心。可以从候选聚类中心中发现已知海洋气候指数的变体和潜在的海洋气候指数。

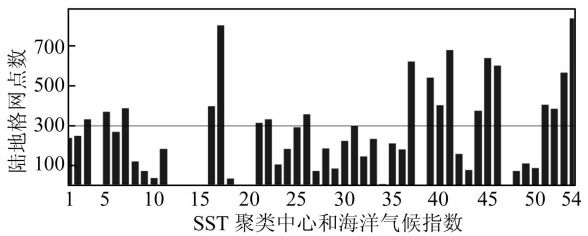

图3 相关性大于0.2的陆地格网点统计直方图

Fig. 3 Statistic Histogram of Land Grid Points Whose Correlation Higher Than 0.2

2.2 挖掘海洋气候指数

2.2.1 挖掘已知海洋气候指数的变体

与海洋气候指数的相关性较大(大于0.4小于0.8)的聚类中心可能与海洋气候指数体现了相似的气候现象,但是这样的聚类中心依然是有价值的,它们可能是已知海洋气候指数的变体。特别是在一些陆地区域,相比海洋气候指数,某些SST聚类中心与陆地气温的相关性更大,并且具有较大的覆盖范围。

为了将SST聚类中心与海洋气候指数进行比较,如果聚类中心与陆地气温的相关性更大,则将相关系数显示为正数;如果海洋气候指数与陆地气温的相关性更大,则将相关系数显示为负数,用颜色条指示不同等级的相关系数值。

如插页Ⅲ彩图5(a)、5(b)、5(c)和5(d)显示了类45与海洋气候指数的比较,右侧的颜色条指示相关系数的大小,红色表示聚类中心与陆地气温的相关性更大,蓝色表示海洋气候指数与陆地气温的相关性更大。插页Ⅲ彩图5(e)用颜色显示的区域为类45的地理位置。可以明显看出,在非洲的大部分地区、西欧、南美洲东部和澳大利亚南部地区,类45与陆地气温的相关性比海洋气候指数与陆地气温的相关性更大。插页Ⅲ彩图5(f)为类45的SST聚类中心的时间序列曲线图,将其与基于海表温度异常的海洋气候指数的曲线图进行比较,可以看到,聚类中心45反映了厄尔尼诺指数所体现的几次典型的海表温度异常,同时还反映了1987~1988年的一次海表温度的陡然增温,这是已知海洋气候指数所没有体现的。

因此,类45可以作为已知海洋气候指数的变体,在研究非洲地区、西欧、南美洲东部和澳大利亚南部海岸等地区的气候变化时,可以采用类45对应的SST聚类中心的时间序列作为厄尔尼诺指数的变体,研究海陆气候变化的关系。

2.2.2 挖掘潜在的海洋气候指数

与海洋气候指数的相关性小于0.4的SST聚类中心可能代表新的海洋气候现象,从这些聚类中心中可能发现潜在的海洋气候指数。插页Ⅳ彩图6为类26与海洋气候指数的比较,可以看到,在非洲大部分地区、印度、中国西南部、南美洲的少部分地区,类26的影响比较显著,相关性胜过几个海洋气候指数。插页Ⅳ彩图6(d)为类26的地理位置,彩图6(e)为类26的SST聚类中心的时间序列曲线图,将其与基于海表温度异常的海洋气候指数的时间序列曲线图进行比较,可以看到,聚类中心26反映了1986年的一次海表温度的陡然下降、1999年海表温度的一次陡然上升以及2001~2002年海表温度的一次陡然下降,这是已知海洋气候指数所没有体现的。类似地,插页Ⅳ彩图7、彩图8分别显示了类41、类37与海洋气候指数的比较。这些海表温度聚类模式可以视为潜在的海洋气候指数。

通过以上分析,从海表温度聚类模式中挖掘出了已知海洋气候指数的变体以及潜在的基于海表温度异常的海洋气候指数。特别是在研究非洲地区、南美洲东北部、澳大利亚北部、东南亚、印度尼西亚、印度、俄罗斯东部、北美洲南部和欧洲的某些国家的气候变化时,可以考虑使用这些SST聚类中心的时间序列作为潜在的海洋气候指数研究海陆气候的变化关系。

3 结 语

本文使用模糊聚类的方法对全球海表温度进

行了聚类分析,已知海洋气候指数得到了重现,验证了模糊聚类的有效性。从模糊聚类得到的聚类模式中,在前人研究的基础上挖掘出了新的可以作为潜在的海洋气候指数的 SST 聚类中心,这些聚类中心与陆地气温的相关性较好,可以对较大范围陆地区域的气候产生影响。本文使用模糊聚类方法挖掘出了新的海洋气候指数,但是仍存在许多不足,如使用二型模糊 C 均值聚类方法对全球海表温度进行聚类的算法效率较低。本文使用的是长时间序列的月平均数据研究大尺度的全球气候变化,今后可以在此基础上使用高分辨率的海表温度数据或其他变量,综合多种陆地气候变量进行分析,并可考虑时滞效应,使用本文得到的 SST 聚类中心和潜在的海洋气候指数研究局部地区的气候变化。

参 考 文 献

[1] 李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论和方法[J]. 武汉大学学报·信息科学版,2002,27(3):221-233

[2] Steinbach M, Tan Pangning, Kumar V,et al. Clustering Earth Science Data,Goals, Issues and Results [C]. KDD Workshop on Mining Scientific Datasets,San Francisco, California, USA,2001

[3] Steinbach M,Tan Pangning , Kumar V, et al. Discovery of Climate Indices Using Clustering[C]. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2003

[4] Sousa F M, Nascimento S, Casimiro H, et al. Identification of Upwelling Areas on Sea Surface Temperature Images Using Fuzzy Clustering [J]. Remote Sensing of Environment, 2008, 112 (6): 2 817-2 823

[5] Qin Kun, Kong Lingqiao, Liu Y, et al. Sea Surface Temperature Clustering Based on Type-2 Fuzzy Theory[C]. The 18th International Conference on Geoinformatics, Beijing,2010

[6] Qin Kun, Wu M R, Kong Lingqiao, et al. Spatio-temporal Data Clustering Based on Type-2 Fuzzy Sets and Cloud Models[C]. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, Hawaii, USA,2010

[7] Steinbach M, Tan Pangning, Kumar V, et al . Data Mining for Discovery of Ocean Climate Indices [C]. KDD Workshop on Temporal Data Mining, Edmonton, Alberta, Canada, 2002

[8] 王新洲. 论空间数据处理与空间数据挖掘[J]. 武汉大学学报·信息科学版,2006,31(1):1-8

[9] Hisdal E. If Then Else Statement and Interval-valued Fuzzy Sets of Higher Type[J]. International Journal of Man-Machine Studies,1981(15):285-455

[10] Hwang C, Rhee F C H. Uncertain Fuzzy Clustering: Interval Type-2 Fuzzy Approach to C-Means [J]. IEEE Transactions on Fuzzy Systems, 2007, 15(1):107-120

[11] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范明, 孟小峰. 北京:机械工业出版社,2007:3, 251,320-322

[12] 冯士筰,李凤歧,李少菁. 海洋科学导论[M]. 北京:高等教育出版社,1999:267-270

第一作者简介:孔令桥,硕士,主要从事空间分析与空间数据挖掘研究。

E-mail:klq1220@126.com

Global SST Data Mining Based on Fuzzy Clustering

KONG Lingqiao[1]　*QIN Kun*[1]　*LONG Tengfei*[2]

(1　School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(2　Center for Earth Observation and Digital Earth, Chinese Academy of Sciences, 9 South Dengzhuang Road, Haidian District, Beijing 100094, China)

Abstract: This paper implements global SST clustering analysis using C-means clustering method based on type 2 fuzzy sets, from which the typical clustering patterns of SST anomaly are discovered, and the potential ocean climate indices are discovered from the clustering patterns.

Key words: fuzzy clustering; SST; data mining; ocean climate indices; climate change

About the first author: KONG Lingqiao, master, majors in spatial analysis and spatial data mining.

E-mail: klq1220@126.com