正态云模型相似度计算方法

李海林[1],郭崇慧[1],邱望仁[2]
(1.大连理工大学系统工程研究所,辽宁大连 116024; 2.大连理工大学信息与控制研究中心,辽宁大连 116024)

摘 要: 本文提出了两种正态云模型相似度计算方法,分别通过正态云模型的期望曲线和最大边界曲线来描述正态云模型的总体特征,实现以期望曲线相似程度或最大边界曲线的相似程度对正态云模型相似度的定量表示.它们在一定程度上克服了传统基于特征向量和随机选取云滴的相似度计算带来云模型期望数字特征过于显著、时间复杂度过高和结果不稳定等方面的不足.实验结果表明,本文算法能够更为客观地对正态云模型进行相似度计算,在协同过滤推荐以及时间序列分类中得到了应用并提高了算法的效率.

关键词: 云模型; 相似性度量; 正态分布; 期望曲线; 最大边界曲线
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2011) 11-2561-07

Similarity Measurement between Normal Cloud Models

LI Hai-lin[1], GUO Chong-hui[1], QIU Wang-ren[2]
(1. *Institute of Systems Engineering, Dalian University of Technology, Dalian, Liaoning* 116024, *China*;
2. *School of Electronic and Information Engineering, Dalian University of Technology, Dalian, Liaoning* 116024, *China*)

Abstract: We proposed two methods to measure the similarity of normal cloud models. One uses the expectation curves to reflect the overall feature of cloud models and to calculate the similarity by the expectation curves. The other uses the maximum boundary curve to compute the similarity between different clouds. The two methods can obtain a qualitative result, which overcomes the traditional deficiencies of the high time complexity, unstable result and excessively remarkable expectation character. The experimental results demonstrate that our methods can calculate the similarity of cloud models objectively and improve the efficiency of the algorithms in collaborative filtering recommendation and time series classification.

Key words: cloud model; similarity measurement; normal distribution; expectation curve; maximum boundary curve

1 引言

自然语言是人类智慧的结晶,在人工智能中具有重要的地位,是通过语言值来表示概念,这些概念通常具有不确定性.以往研究不确定性的方法有很多,如概率论、模糊集理论、粗糙集理论等,但利用这些方法来研究概念的不确定性时尚存在一定的局限性.特别地,在研究自然语言的模糊性和随机性时,没有很好地将两者联系起来.云理论就是同时从这两者的角度来研究自然语言的不确定性,建立云模型来实现定性概念与定量表示之间的转化[1].

近年来,在数据挖掘领域中,云模型不仅运用于挖掘过程的不确定性表示,而且为挖掘结果的表示提供了符合人类思维习惯的定性分析方法.通常情况下,数据挖掘任务中的定量数据可以通过云模型来实现定性概念转换,同时建立在定性概念基础之上的数据挖掘任务需要进行相似性计算,例如分类、聚类、相似性搜索等.特别地,在股票时间序列数据挖掘中,云模型可以对时间序列数据进行分段概念表示,需要利用云模型相似性计算方法来度量概念之间的距离,以便在挖掘过程中发现潜在的序列模式和其它信息.因此,在云数据挖掘应用领域中,云模型相似度计算方法的优劣直接影响到数据挖掘算法的效率.传统云模型相似度计算方法是基于特征向量或随机选取云滴进行相似度比较,例如,文[2]提出的相似云及其度量方法就是一种通过随机取若干个云滴,计算这些云滴的距离值来表示云模型间的相似性;文献[3]提出基于云模型的协同过滤推荐算法,将云模型的数字特征当作向量,并且利用夹角余弦来衡量云模型之间的相似度问题.前者虽然能够随机地表示云模型的相似度,但选取云滴、对云滴的排序以及云滴的组合所消耗的时间将不利于大规模数据;后者把数字特征作为向量直接利用夹角余弦来得到云模型的相似度,虽然在协同过滤算法中取得了较好的效果,但很多情况下云模型的数字特征中的期望值远远大于熵和超熵,使得夹角余弦的度量容易忽视熵和超熵两个数字特征的

收稿日期:2011-01-11;修回日期:2011-03-29
基金项目:国家自然科学基金(No.10571018,No.70871015);中央高校基本科研业务专项资金(No.DUT11SX04)

作用.因此,本文分别提出基于期望曲线的云模型相似度计算方法和基于最大边界曲线的云模型相似度计算方法,它们在一定程度上克服了传统方法的不足.实验结果表明,这两种方法能更为客观地衡量云模型的相似性,进而有效地提高了数据挖掘算法的效率.

2 云模型

云模型是不确定性人工智能中研究自然语言从定性概念到定量表示之间相互转化的一种模型,主要反映了人类知识中概念的模糊性和随机性,为研究不确定性人工智能提供了新的方法[4~6].它已经在数据挖掘和知识发现[7,8]、信号识别[9]以及决策分析[10]等方面得到了广泛的应用并取得了良好的效果.正态云是一种较为重要和普遍的模型,具有良好的数学性质并且现实世界中许多现象都服从或近似服从正态分布,因此,它具有一定的普适性.

定义 1 设 U 是一个用精确数值表示的定量论域,C 是 U 上的定性概念,若定量值 $x\in U$,且 x 是定性概念 C 的一次随机实现,x 对 C 的确定度为 $\mu_C(x)\in[0,1]$是具有稳定倾向的随机数,则 x 在论域 U 上的分布称为云,每个 x 称为一个云滴.

云是由若干云滴组成,云滴是某个定性概念的一次随机实现,多次产生的云滴可以综合反映这个定性概念的整体特征.某个概念的整体特征可以用云的 3 个数字特征来表示,即期望 Ex、熵 En 和超熵 He,因此,云模型特征可以由这三个数字特征组成的向量来描述.

定义 2 由三个参数(Ex,En,He)来表示云的数字特征的模型,称为云模型(Ex,En,He).其中,云的期望 Ex 表示云滴在论域空间分布的期望值,即最能够代表定性概念的点;云的熵 En 表示定性概念的不确定性度量,可以用来描述云的跨度,反映云滴的离散程度;超熵 He 是熵的不确定性度量,可以用来描述云的厚度,如图 1 所示.

定义 3 若随机变量 x 满足:$x\sim N(Ex,En'^2)$,其中 $En'\sim N(En,He^2)$,对定性概念 C 的确定度满足:

$$\mu_C(x)=e^{-\frac{(x-Ex)^2}{2En'^2}} \tag{1}$$

则 x 在论域 U 上的分布称为正态云.

正态云是一种定性概念与定量表示之间的不确定转化模型.正向正态云发生器是实现从定性概念到定量数据的映射;相反,从定量表示到定性概念的映射可以由逆向正态云发生器来实现.

由正向正态云发生器产生的云滴中,各个云滴对特定概念的贡献不同.与正态分布类似,对于定性概念有重要贡献的云滴主要落在区间$[Ex-3En,Ex+3En]$中.位于$[Ex-3En,Ex+3En]$之外的云滴元素称为小概率事件,忽略并不影响它的整体特征,这就是正向正态云的 $3En$ 规则[1,11],相当于正态分布的 3σ 规则,如图 1 所示.

由图 1 可知,正态云有明显的几何特征,通常可以借助回归曲线和主曲线来研究其特性.这两种曲线分别从垂直方向的期望和正交方向的期望来反映云的整体特征[1],但由于它们的解析式难于求出,只能通过线性逼近的方法近似求得.然而,期望曲线是从水平方向来研究云模型整体特征,通过正态云的定义可以推出期望曲线的解析式.

定义 4 若随机变量 x 满足:$x\sim N(Ex,En'^2)$,其中 $En'\sim N(En,He^2)$且 $En\neq 0$,则

$$y=e^{-\frac{(x-Ex)^2}{2En^2}} \tag{2}$$

称为正态云的期望曲线.

从图 2 可以看出,用期望曲线方法可以很好地反映正态云的重要几何特征,所有的云滴都围绕正态云期望曲线这条“骨架”的附近随机波动.

3 正态云模型相似度

3.1 基于期望曲线的正态云模型相似度计算方法

由于具有解析式的正态云期望曲线能够方便有效地描述正态云的总体特征,因此,可以借助正态云期望曲线来求解云模型的相似度.通过求解两个云模型的期望曲线相交重叠部分的面积 S 来表示两个云模型的相似程度,如图 3 所示,阴影部分的面积反映了两个云模型的相似程度.

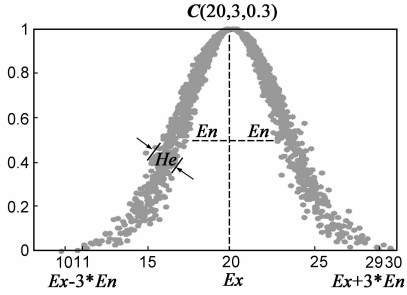


图1 正态云模型$C(20,3,0.3)$

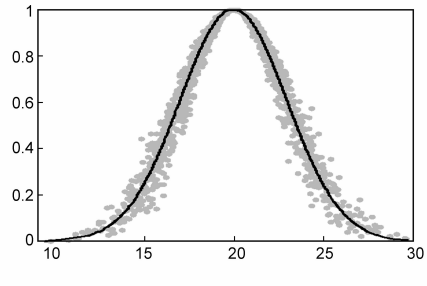


图2 正态云期望曲线

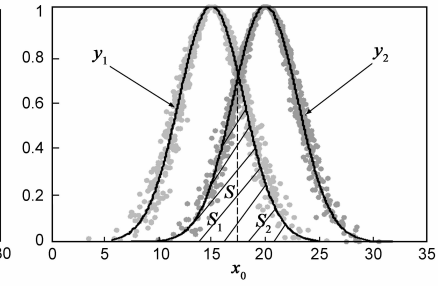


图3 云模型$C(15,3,0.35)$和$C(20,3,0.3)$的相似度面积S

通常情况下,已知两条期望曲线的解析式,可以通过积分方法来求面积 S.如果两条期望曲线相交点的横坐标为 x_0,那么

$$S=\int_{-\infty}^{x_0} y_2(x)\mathrm{d}x+\int_{x_0}^{\infty} y_1(x)\mathrm{d}x \tag{3}$$

其中,$y_1(x)$和 $y_2(x)$分别为正态云模型 C_1 和 C_2 的期望曲线方程.但由于式(2)是一个不可积函数,故只能通过数值逼近方法得到式(3)的近似解.然而,逼近求解方法相当耗费时间,对于大量云模型彼此之间的相似度计算是不可行的.因此,利用一般的数值积分求解方法不适合云模型相似度的计算.

从期望曲线解析式(2)易知,该曲线类似于正态分布的概率密度函数,对其进行变形可得

$$y=\sqrt{2\pi}En\frac{1}{\sqrt{2\pi}En}e^{-\frac{(x-Ex)^2}{2En^2}}=\sqrt{2\pi}En\,f(x) \tag{4}$$

其中 $f(x)$就是正态分布的概率密度函数.利用正态分布的相关性质对重叠部分的面积进行求解,即

$$\begin{aligned}S&=\int_{-\infty}^{x_0} y_2(x)\mathrm{d}x+\int_{x_0}^{\infty} y_1(x)\mathrm{d}x\\&=\sqrt{2\pi}En_2\int_{-\infty}^{x_0} f_2(x)\mathrm{d}x+\sqrt{2\pi}En_1\int_{x_0}^{\infty} f_1(x)\mathrm{d}x\end{aligned} \tag{5}$$

由一般正态分布和标准正态分布的关系,得到

$$\begin{aligned}S&=\sqrt{2\pi}En_2\int_{-\infty}^{x_0} f_2(x)\mathrm{d}x+\sqrt{2\pi}En_1\int_{x_0}^{\infty} f_1(x)\mathrm{d}x\\&=\sqrt{2\pi}En_2\int_{-\infty}^{z_2}\phi(x)\mathrm{d}x+\sqrt{2\pi}En_1\Big(1-\int_{-\infty}^{z_1}\phi(x)\mathrm{d}x\Big)\end{aligned} \tag{6}$$

其中 $z_1=\dfrac{x_0-Ex_1}{En_1}$,$z_2=\dfrac{x_0-Ex_2}{En_2}$,$\phi(x)$为标准正态分布概率密度函数.如果知道 x_0 值,则可以得到 z_1 和 z_2,再结合标准正态分布表,便可求得相交面积 S.

由期望曲线解析式(2)易知两个云模型的期望曲线分别为:

$$\begin{cases}y_1(x)=e^{-\frac{(x-Ex_1)^2}{2En_1^2}}\\ y_2(x)=e^{-\frac{(x-Ex_2)^2}{2En_2^2}}\end{cases} \tag{7}$$

若曲线相交,则有 $y_1(x)=y_2(x)$,即 $|z_1|=|z_2|$,解得:

$$\begin{cases}x_0^{(1)}=\dfrac{Ex_2En_1-Ex_1En_2}{En_1-En_2}\\ x_0^{(2)}=\dfrac{Ex_1En_2+Ex_2En_1}{En_1+En_2}\end{cases} \tag{8}$$

由正向正态云的 $3En$ 规则可知,有 99.74% 的云滴或元素会落在区间$[Ex-3En,Ex+3En]$,在计算正态云相似度时,只考虑该区间上的云滴分布便可.在两个云模型中,不妨设 $Ex_1\leqslant Ex_2$,则两云模型的期望曲线的两个交点 $x_0^{(1)}$ 和 $x_0^{(2)}$ 的分布情况存在以下三种可能.

(1)若 $x_0^{(1)}$ 和 $x_0^{(2)}$ 同时落在区间$[Ex_2-3En_2,Ex_1+3En_1]$外,说明两个交点之间的云滴可以忽略,故相交面积可以视为 0,即 $S=0$.

(2)若 $x_0^{(1)}$ 和 $x_0^{(2)}$ 有一个点落在区间$[Ex_2-3En_2,Ex_1+3En_1]$中,相交情况如图 3 所示,则相交面积由两部分构成,即 $S=S_1+S_2$.

(3)若 $x_0^{(1)}$ 和 $x_0^{(2)}$ 同时落在区间$[Ex_2-3En_2,Ex_1+3En_1]$中,相交情况如图 4 和图 5 所示,则相交面积由三部分构成,即 $S=S_1+S_2+S_3$.

验证 $x_0^{(1)}$ 和 $x_0^{(2)}$ 是否满足正向正态云的 $3En$ 规则并且根据它们的分布情况来判断属于哪种可能.若 x_0 的分布属于第一种情况,则 $S=0$;若 x_0 的分布属于第二种情况,则可以根据式(6)计算出面积 S.若 x_0 分布属于第三种情况,则按图 3 情况的思想分三步求解 S,即 $S=S_1+S_2+S_3$.

在两个云模型 C_1 和 C_2 中,不妨设 $Ex_1\leqslant Ex_2$.对于第三种情况,存在另外两种可能性.

(1)若 $En_1\leqslant En_2$,则面积 S 中,S_1 和 S_3 由云模型 C_1 的期望曲线 y_1 构成,S_2 由云模型 C_2 的期曲望线 y_2 构成,如图 4 所示.

(2)若 $En_1>En_2$,则面积 S 中,S_1 和 S_3 由云模型 C_2 的期望曲线 y_2 构成,S_2 由云模型 C_1 的期望曲线 y_1 构成,如图 5 所示.

根据图 3 情况的分析思想,可以对这两种情况分步求解面积 S.由于可能性(1)和(2)的面积求解方法类似,因此下面给出第(1)种可能性的面积求解方法.

如图 4 所示,两个云模型的熵满足 $En_1\leqslant En_2$,期望曲线的交点分别为 $x_0^{(1)}$ 和 $x_0^{(2)}$,它们的分布满足第三种情况且有 $x_0^{(1)}\leqslant x_0^{(2)}$,则可以推导出面积 S_1、S_2 和 S_3 的求解公式.

面积 S_1 由云模型 C_1 的期望曲线 y_1 构成,有

$$S_1=\sqrt{2\pi}En_1\int_{-\infty}^{x_0^{(1)}} f_1(x)\mathrm{d}x=\sqrt{2\pi}En_1\int_{-\infty}^{z_1^{(1)}}\phi(x)\mathrm{d}x \tag{9}$$

其中 $z_i^{(j)}=\dfrac{x_0^{(j)}-Ex_i}{En_i}$.

面积 S_2 由云模型 C_2 的期望曲线 y_2 构成,有

$$\begin{aligned}S_2&=\sqrt{2\pi}En_2\int_{x_0^{(1)}}^{x_0^{(2)}} f_2(x)\mathrm{d}x\\&=\sqrt{2\pi}En_2\Big(\int_{-\infty}^{x_0^{(2)}} f_2(x)\mathrm{d}x-\int_{-\infty}^{x_0^{(1)}} f_2(x)\mathrm{d}x\Big)\\&=\sqrt{2\pi}En_2\Big(\int_{-\infty}^{z_2^{(2)}}\phi(x)\mathrm{d}x-\int_{-\infty}^{z_2^{(1)}}\phi(x)\mathrm{d}x\Big)\end{aligned} \tag{10}$$

面积 S_3 由云模型 C_1 的期望曲线 y_1 构成,有

$$S_3 = \sqrt{2\pi}En_1\int_{x_0^{(2)}}^{\infty} f_1(x)\mathrm{d}x$$
$$= \sqrt{2\pi}En_1\left(1-\int_{-\infty}^{x_0^{(2)}} f_1(x)\mathrm{d}x\right)$$
$$= \sqrt{2\pi}En_1\left(1-\int_{-\infty}^{z_1^{(2)}} \phi(x)\mathrm{d}x\right) \quad (11)$$

通过查询标准正态分布表,便可以快速解出两个云模型重叠的三部分面积 $S = S_1 + S_2 + S_3$.

为了对不同云模型的相似度进行比较,必须对面积 S 做标准化处理,最终得到基于期望曲线的云模型相似度(Expectation based Cloud Model,ECM),

$$ECM(C_1, C_2) = \frac{2S}{\sqrt{2\pi}(En_1 + En_2)} \in [0,1] \quad (12)$$

其中$\sqrt{2\pi}En_1$和$\sqrt{2\pi}En_2$分别表示两个正态云模型的期望曲线与横坐标之间形成的面积.

算法 1 正态云模型相似度算法

输入:两个正态云模型 $C_1(Ex_1, En_1, He_1)$和 $C_2(Ex_2, En_2, He_2)$.

输出:正态云模型相似度 $ECM(C_1, C_2)$.

Step 1 不妨设 $Ex_1 \leqslant Ex_2$ 且初始设置 $S = 0$.通过公式(8)求解 $x_0^{(1)}$ 与 $x_0^{(2)}$ 值,不妨设 $x_0^{(1)} \leqslant x_0^{(2)}$.

Step 2 若 $x_0^{(1)} \leqslant \min(Ex_1 - 3En_1, Ex_2 - 3En_2)$且 $x_0^{(2)} \geqslant \max(Ex_1 + 3En_1, Ex_2 + 3En_2)$时,则 $ECM(C_1, C_2) = 0$ 并且程序停止;否则,执行下一步.

Step 3 若 $x_0^{(1)} \geqslant \max(Ex_1 - 3En_1, Ex_2 - 3En_2)$且 $x_0^{(2)} \leqslant \min(Ex_1 + 3En_1, Ex_2 + 3En_2)$时,则根据两个云模型的熵($En_1$ 和 En_2)大小情况来求解面积 $S = S_1 + S_2 + S_3$;否则,执行下一步.

Step 4 在其它情况下,$x_0^{(1)}$ 或 $x_0^{(2)}$ 会落在区间$[Ex_2 - 3En_2, Ex_1 + 3En_1]$中,即 $S = S_1 + S_2$.

Step 5 将 S 代入式(12),计算出 $ECM(C_1, C_2)$.

综上所述,正态云模型相似度 $ECM(C_1, C_2)$计算过程如算法 1.

3.2 基于最大边界曲线的正态云相似度计算方法

基于期望曲线的正态云相似度计算方法,主要是通过期望曲线来描述不同云模型之间的相似度.期望曲线能够很好地反映云模型的整体特征,是一种从整体几何特征的角度来研究正态云模型的相似度,进而可以忽略云模型中超熵的描述.然而,很多情况下,要从局部的角度来研究云模型的相似度,因此,下面提出一种基于最大边界曲线的正态云相似度计算方法(Maximum boundary based Cloud Model,MCM).它能够让云模型的三个数字特征都参与相似度计算,实现云模型局部特征的相似度计算.

利用正态云定义以及正向正态云模型的 $3En$ 规则,正态云模型的最大边界曲线解析式可定义为:

$$y = e^{-\frac{(x-Ex)^2}{2(3He+En)^2}} \quad (13)$$

从图 6 中可以发现,几乎所有的云滴都在这条最大边界曲线之下,这是由正态分布的 3σ 规则所决定的.最大边界曲线是一种从最大云滴值这个局部性视角来研究云模型几何特性的方法.

式(13)与式(2)很相似,都是不可积的解析式,因此可以按照上一节方法来研究基于最大边界曲线的正态云相似度计算方法.令 $en = 3He + En$,则式(13)将变成

$$y = e^{-\frac{(x-Ex)^2}{2en^2}} \quad (14)$$

得到与期望曲线方法相似的最大边界曲线解析式,同样,可以通过类似基于期望曲线的云模型相似度求解方法来得到两个云模型之间最大边界曲线相重叠部分的面积,最终得到两个正态云模型的相似度.

从式(2)和式(13)并且结合图 2 和图 6 中不难发现,基于正态云期望曲线的相似性计算方法(ECM)考虑了云模型的前两个数字特征,即期望 Ex 和熵 En,从云模型的期望位置和跨度的角度来比较正态云的相似性;基于最大边界曲线的正态云相似性计算方法(MCM)同时考虑了正态云模型的三个数字特征,即期望 Ex、熵 En 和超熵 He,增加了对正态云厚度的理解,从更微观(局部)的角度来比较正态云的相似性.

4 实验及分析

首先通过仿真实例来分别对 ECM 和 MCM 进行数值实验,并分析比较 ECM 算法、MCM 算法、文[2]提出的(Similar Cloud Measurement,SCM)算法和文[3]提出的(LIkeness comparing method based on Cloud Model,LICM)算法的结果和联系.其次,利用电影评价的真实数据集,通过协同过滤推荐算法分别对 ECM、MCM 和 LICM 三种

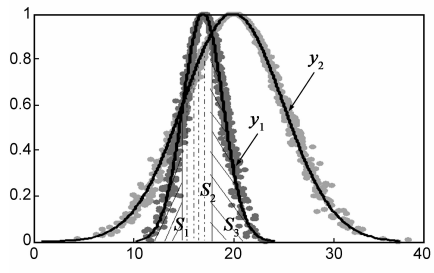


图4 云模型$C_1(17,2,0.25)$和$C(20,5,0.3)$的相似度面积$S=S_1+S_2+S_3$,$En_1 \leqslant En_2$

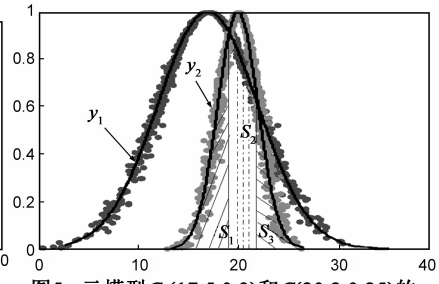


图5 云模型$C_1(17,5,0.3)$和$C(20,2,0.25)$的相似度面积$S=S_1+S_2+S_3$,$En_1 > En_2$

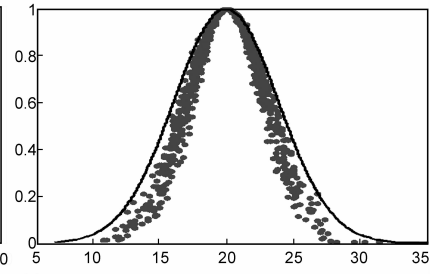


图6 云模型的最大边界曲线

算法进行实验并分析比较,进而验证 ECM 和 MCM 的可行性和有效性.最后,为了进一步验证 ECM 和 MCM 的效率,让这四种方法对高维时间序列数据进行分类实验,从算法实验结果的精度和程序运行速度进行比较分析,进而说明本文提出的两种算法的效率.

4.1 仿真实验

为了更好地说明本文提出的两种云模型相似度计算方法和 SCM 算法、LICM 算法之间的差异性,分别利用文[2]和文[3]中示例数据进行数据实验,并且分析比较它们之间的实验结果.

在文[2]实验中选取的两组云模型分别为: $G_1=[C_1,C_2]$和 $G_2=[C_2,C_3]$,其中 $C_1=(3,3.123,2.05)$、$C_2=(2,3,1)$和 $C_3=(1.585,3.556,1.358)$.SCM 算法的计算结果为 $Distance(G_1)=0.0428$ 和 $Distance(G_2)=0.029$,前者大于距离阈值 0.03,后者小于距离阈值 0.03,故 C_1 与 C_2 不相似,C_2 与 C_3 相似.

利用 SCM 对$[C_1,C_2,C_3]$进行 50 次云模型距离度量实验,取它们的平均值作为距离度量实验结果.同时,分别运用本文提出的 ECM 和 MCM 对它们进行相似度计算,实验结果如表 1 所示.

表 1　SCM、ECM 与 MCM 算法度量云模型的距离相似性

	SCM(距离)			ECM			MCM		
	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
C_1	0.0000	**0.0373**	0.0386	1.0000	**0.8728**	0.8336	1.0000	**0.7821**	0.8983
C_2	0.0373	0.0000	**0.0094**	0.8728	1.0000	**0.9138**	0.7821	1.0000	**0.8800**
C_3	0.0386	0.0094	0.0000	0.8336	0.9138	1.0000	0.8983	0.8800	1.0000

可以发现,ECM 和 MCM 可以得出与 SCM 一样的结论,即 C_2 和 C_3 之间相似性大于 C_1 与 C_2 之间的相似性,并且由于 SCM 随机选取云滴会引起最终相似度结果不稳定,而 ECM 和 MCM 是云模型期望曲线和最大边界曲线的度量方法,这两条固定的曲线决定了新算法计算结果的稳定性.然而,MCM 的实验结果还表明 C_1 和 C_3 最相似,这与 ECM 和 SCM 不一致,其主要原因是过分大的超熵 He 对正态云模型的最大边界曲线影响较大,进而影响云模型相似性计算.在一般情况下,超熵 He 的值会小于 1,不容易出现实验中过大超熵值的现象.

文[3]的 LICM 算法是一种把云模型数字特征看作向量元素,利用夹角余弦进行相似度求解的方法,该方法在协同过滤推荐领域中得到了较好的应用.对于 4 个云模型,$C_1=[1.5000,0.62666,0.3390]$、$C_2=[4.6000,0.60159,0.30862]$、$C_3=[4.4000,0.75199,0.27676]$和 $C_4=[1.6000,0.60159,0.30862]$.分别利用 LICM、ECM 和 MCM 进行计算,其结果如表 2 所示.可以发现,其结论与 LICM 一致,即 C_1 和 C_4 为一类,C_2 和 C_3 为一类,并且 ECM 和 MCM 更能体现不同类中云模型之间的差异性.例如,针对不同类的 C_1 与 C_2,利用 LICM 计算它们相似度为 0.96,而 ECM 和 MCM 计算它们的相似度分别为 0.01 和 0.33.显然,本文提出的两种算法更能体现不同类别中云模型的差异性.

表 2　LICM、ECM 与 MCM 算法计算云模型的相似性

	LICM				ECM				MCM			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
C_1	1.00	0.96	0.97	**0.99**	1.00	0.01	0.04	**0.94**	1.00	0.33	0.37	**0.96**
C_2	0.96	1.00	**0.99**	0.97	0.01	1.00	**0.86**	0.01	0.33	1.00	**0.95**	0.38
C_3	0.97	**0.99**	1.00	0.98	0.04	**0.86**	1.00	0.04	0.37	**0.95**	1.00	0.37
C_4	**0.99**	0.97	0.98	1.00	**0.94**	0.01	0.04	1.00	**0.96**	0.33	0.37	1.00

4.2 协同过滤推荐实验

为了验证 ECM 和 MCM 的可行性和有效性,使用 MoiveLens 站点提供的数据集[12],选取 1997-9-19 到 1998-5-22 的数据集进行电影评价协同推荐.该数据集总共有 100000 条记录,每条记录包括 4 个属性,分别是用户标识、电影标识、用户对电影的评价和时间标识.该数据集记录了 943 个用户对 1682 个影片的评价记录,且分为训练集(80000 个记录)和测试集(20000 个记录),文[3]已经证实 LICM 相似性度量方法在协同过滤推荐算法中要优于余弦相似性、修正余弦相似性和 BP_CF (back propagation-collaborative filtering)[13],因此,本次实验只进行 LICM、ECM 和 MCM 之间的比较分析.

同样,利用平均绝对偏差(MAE)[14,15]来说明预测的准确性,即 MAE 越大,预测越不准确,推荐质量就越差.通过实验计算得到三种基于云模型的相似度计算方法对协同过滤推荐算法的结果,如图 7 所示.

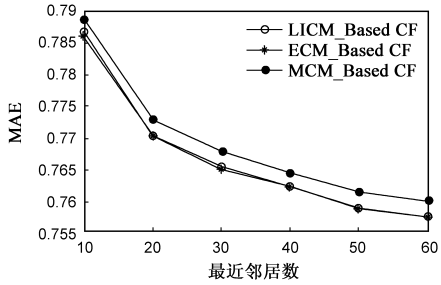


图7　ECM、LICM和MCM三种方法的MAE随最近邻居数的变化

容易发现,基于 ECM 的协同过滤推荐质量大体与 LICM 保持一致,甚至在有些最近邻居数,ECM 的 MAE 要小于 LICM.因此可以说,基于 ECM 的协同过滤推荐质量总体上要略优于 LICM.虽然 MCM 的 MAE 大于前两者,但差值在 0.01 之内,趋势与前两者保持一致.同时通过比较文[3]的实验,MCM 的 MAE 都会小于余弦相似性、修正余弦相似性以及 BP_CF,并且总体趋势与 LICM 保持一致.因此,该实验结果验证了 ECM 和 MCM 的可行性和有效性.

4.3 时间序列分类实验及时间代价分析

由于时间序列具有高维性特点,能够很好地检验

分类算法的可行性和分类结果的准确度.然而,分类结果的准确度除了依靠算法本身的分类能力外,还要取决于分类算法过程中所运用的相似度度量方法.

本次实验使用 UCI 中的时间序列数据集(synthetic control chart dataset)[16],该数据集共有 6 类数据,每组数据各含有 100 个长度为 60 的时间序列.在本次实验中.分别取每一类的最后 11 个时间序列作为测试集,其余时间序列作为训练集,最终共有 534 个时间序列组成训练集和 66 个时间序列构成测试集.

为了验证 ECM 和 MCM 的准确性和效率,统一采用最近邻分类算法[17,18]分别对 ECM、MCM、LICM 和 SCM 等四种算法进行分类实验.最终通过分类的错误率来评价四种算法的分类准确能力.同时,为了提高时间序列挖掘的效率,通常对时间序列进行降维处理[18,19].因此,将每个时间序列进行平均分段分类实验,其分段数目(即降维后的维数)分别 2、3、5、10、15 和 20.例如,每个时间序列按时间顺序平均分成 2 段,每一段 1 个云模型来表示,那么整个时间序列将可以用 2 个云模型来表示.最终四种相似度计算方法分别在不同分段数(维数)下的分类结果如图 8 所示.

从图 8 可以看出,ECM 的分类错误率明显低于 LICM,甚至分类错误率接近于 0.虽然 SCM 的分类错误率低于 LICM,但也不如 ECM。MCM 却介于 SCM 和 LICM 之间,但更趋近于 SCM 的分类精度,甚至在有些维度下,MCM 与 SCM 有相同的分类精度。实事上,由于 LICM 算法将云模型数字特征看成特征向量,运用夹角余弦将会因期望值相对于熵和越熵过大而造成期望数字特征过于显著,容易忽视其它两个数字特征的作用.另外,ECM 和 MCM 随着降维数的变化能够保持稳定的分类精度,说明这两种算法具有很好的伸缩性和鲁棒性.

在分类准确性方面,虽然 SCM 比较接近 ECM,甚至当维度为 20 时,出现两者分类准确度相同的结果.但从时间复杂度的角度出发,如图 9 所示,除 SCM 之外,其余三种云模型相似性度量算法所消耗的单位时间相差不大,它们时间复杂度相同.然而,SCM 算法的时间复杂度最大,不利于时间序列的分类.其原因在于:在 SCM 算法过程中,为了较精确地度量云模型相似度,不仅需要正向正态云发生器产生足够多的云滴,而且还需要对这些云滴进行组合排序,这些操作都是比较消耗时间的过程.因此,SCM 算法不适合用于诸如时间序列高维大规模数据的相似性比较.

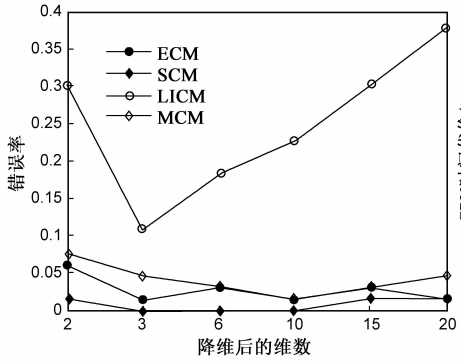
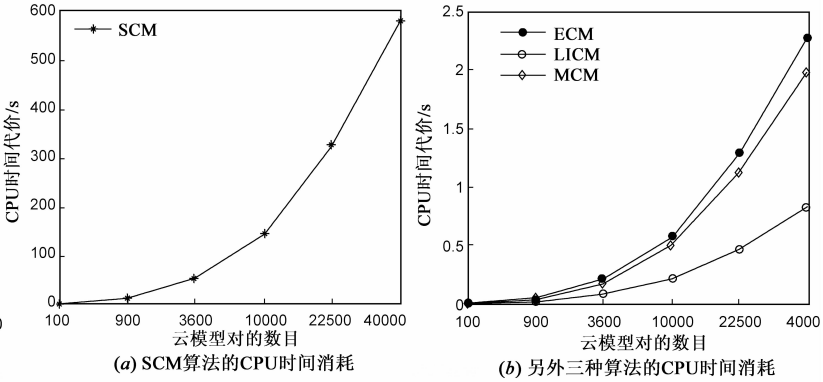


图8 四种相似度计算方法在不同维度下的分类错误率



(*a*) SCM算法的CPU时间消耗 (*b*) 另外三种算法的CPU时间消耗

图9 四种云模型相似度计算方法的CPU时间代价

通过上面实验可以知道,在四种算法中,ECM 不仅在时间序列分类结果的精度上取得优势,而且在时间复杂度上也几乎跟 LICM 持平.因此,ECM 是一种快速有效的云模型相似度计算方法,而 MCM 也是一种较为快速有效的云模型相似度计算方法.

5 总结

本文提出了 ECM 和 MCM 两种新的正态云模型相似度计算方法.ECM 是一种基于正态云模型期望曲线的相似度计算方法,它是利用云模型的"骨架"并且结合查询标准正态分布表来快速计算出正态云之间的相似度.而 MCM 是基于最大边界曲线的正态正态云相似度计算方法,它综合利用了云的三个数字特征,从最大边界这个局部视角来研究相似性的定量数值.同时,不仅从数学性质的角度对 ECM 和 MCM 的正态云模型相似度算法进行了分析和描述,而且还通过实验验证了这两种方法进行描述正态云模型相似程度的可行性和有效性,并将它运用于协同过滤推荐和时间序列数据分类中,取得了良好的效果.然而,云是一种具有定性描述数据能力的模型,如何运用云相似度计算方法去定性分析数据之间的关系是今后研究的一个方向.

参考文献

[1] 李德毅,杜 ▮ .不确定性人工智能[M].国防工业出版社,2005.

[2] 张勇,赵东宁,李德毅.相似云及其度量分析方法[J].信

息与控制,2004,33(2):130 – 132.
Zhang Yong, Zhao Dongning, Li Deyi. The similar cloud and the measurement method[J]. Information and Control, 2004, 33(2):130 – 132. (in Chinese)

[3] 张光卫,李德毅,李鹏,等.基于云模型协同过滤推荐算法[J].软件学报,2007,18(10):2403 – 2411.
Zhang Guangwei, Li Deyi, Li Peng, et al. A collaborative filtering recommendation algorithm based on cloud model[J]. Journal of Software, 2007, 18(10):2403 – 2411. (in Chinese)

[4] 李德毅.知识表示中的不确定性[J].中国工程科学,2000,2(10):73 – 79.
Li Deyi. Uncertainty in knowledge representation[J]. Engineering Science, 2000, 2(10):73 – 79. (in Chinese)

[5] Li D Y, Han J W, Shi X M, et al. Knowledge representation and discovery based on linguistic atoms[J]. Knowledge-based Systems, 1998(10):431 – 440.

[6] Li D Y. Knowledge representation in KDD based on linguistic atoms[J]. Journal of Computer Science and Technology, 1997, 12(6):481 – 496.

[7] 戴朝华,朱云芳,陈维荣,等.云遗传算法及其应用[J].电子学报,2007,35(7):1421 – 1424.
Dai Chaohua, Zhu Yunfang, Chen Weirong, et al. Cloud model based genetic algorithm and its applications[J]. Acta Electronica Sinica, 2007, 35(7):1421 – 1424. (in Chinese)

[8] 刘禹,李德毅,张光卫,等.云模型雾化特性及在进化算法中的应用[J].电子学报,2009,37(8):1651 – 1658.
Liu Yu, Li Deyi, Zhang Guangwei, et al. Atomized feature in cloud based evolutionary algorithm[J]. Acta Electronica Sinica, 2009, 37(8):1651 – 1658. (in Chinese)

[9] 海军,柳征,姜文利,等.基于云模型和矢量神经网络的辐射源识别方法[J].电子学报,2010,38(12):2797 – 2804.
Hai Jun, Liu Zheng, Jiang Wenli, et al. Approach based on cloud model and vector neural network for emitter identification [J]. Acta Electronica Sinica. 2010, 38(12):2797 – 2804. (in Chinese)

[10] 柳炳祥,李海林,杨丽彬.云决策分析方法[J].控制与决策,2009,24(6):957 – 960.
Liu Bingxiang, Li Hailin, Yang Libin. Cloud decision analysis method[J]. Control and Decision, 2009, 24(6):957 – 960. (in Chinese)

[11] 王晓峰,洪磊.基于云的概念空间模型研究.计算机工程与应用[J].2010,46(20):202 – 206.
WANG Xiaofeng, HONG Lei. Study on concept space model based on the cloud theory Computer[J]. Engineering and Applications, 2010, 46(20):202 – 206. (in Chinese)

[12] MoiveLens[EB/OL]. http://movielens.umn.edu, 1997-9-19/ 1998-5-22.

[13] 张锋,常会友.使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J].计算机研究与发展,2006,43(4):667 – 672.
Zhang Feng, Chang Huiyou. Employing BP neural networks to alleviate the sparsity issue in collaborative filtering recommendation algorithms[J]. Journal of Computer Research and Development, 2006, 43(4):667 – 672. (in Chinese)

[14] 张丙奇.基于领域知识的个性化推荐算法研究[J].计算机工程,2005,31(21):7 – 9.
Zhang Bingqi. A collaborative filtering recommendation algorithm based on domain knowledge[J]. Computer Engineering, 2005, 31(21):7 – 9. (in Chinese)

[15] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[A]. Proceeding of the 10th Intelligent World Wide Web Conference. Hong Kong: ACM Press, 2001. 285 – 295.

[16] Pham D T, Chan A B. Control chart pattern recognition using a new type of self organizing neural network[J]. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 1998, 212(1):115 – 127.

[17] Fuchs E, Gruber T, Pree H, et al. Temporal data mining using shape space representation of time series[J]. Neurocomputing, 2010, 74:379 – 393.

[18] Lin J, Keogh E, Li W, et al. Experiencing SAX: a novel symbolic representation of time series[J]. Data Mining and Knowledge Discovery, 2007, 15:107 – 144.

[19] 闫秋艳,夏士雄.一种无限长时间序列的分段线性拟合算法[J].电子学报,2010,38(2):443 – 448.
Yan Qiuyan, Xia Shixiong. An piecewise linear fitting algorithm for infinite time series[J]. Acta Electronica Sinica, 2010, 38(2):443 – 448. (in Chinese)

作者简介

李海林　男,1982 年生于福建龙岩,博士研究生.研究方向为数据挖掘、人工智能等.
E-mail:hailin@mail.dlut.edu.cn

郭崇慧　男,1973 年生于辽宁丹东,教授,博士生导师,研究方向为数据挖掘与知识发现,决策理论与方法等.