

## 基于云模型重叠度的相似性度量

孙妮妮, 陈泽华, 牛昱光, 阎高伟\*

(太原理工大学 信息工程学院, 太原 030024)

(\*通信作者电子邮箱 yangaowei@tyut.edu.cn)

**摘要:**云模型相似性是用来度量同类概念不同语言值的多个云之间关联程度的方法,相似云及其度量分析方法提出是对云模型理论的扩展。针对目前相似性度量方法中时间复杂度过高和结果不稳定等不足,提出了一种基于云模型重叠度的相似性度量算法。首先,根据云模型期望、熵、超熵三个数字特征,定义两个云模型的位置关系和逻辑关系;其次,利用两个云的位置和形状特性,计算得到它们间的重叠度;最后,结合云模型重叠度与相似度的关系,将云模型的相似性度量转化为相应重叠部分的定量描述。通过对时间序列分类实例的应用,验证了该算法在保证结果稳定性和正确率的前提下,与目前时间消耗较低的云模型相似度计算方法(LICM)相比,计算复杂度降低了50%,表明该算法具有可行性和有效性。

**关键词:**云模型;相似性;重叠度;逻辑关系;度量算法;时间序列

**中图分类号:** TP181 **文献标志码:** A

### Similarity measurement between cloud models based on overlap degree

SUN Nini, CHEN Zehua, NIU Yuguang, YAN Gaowei

(College of Information Engineering, Taiyuan University of Technology, Taiyuan Shanxi 030024, China)

**Abstract:** Similarity measurement of cloud model is a method that is used to measure the correlation between cloud models, which have same concept but different languages. Both similar cloud and its measurement analysis method are the extension of cloud model theory. To overcome the disadvantages of high consumption and low precision of calculation, a similarity measure algorithm based on overlap degree was proposed. Firstly, the position and logical relationships between these two clouds were defined according to three digital features: expected value, entropy and hyper entropy; secondly, the overlap degree of two clouds was calculated by using their location and shape features; finally, combined with overlap degree and similarity, the similarity measurement was converted to quantitative description of the overlapping part. In the time series classification experiments with compared Likeness comparing method based on Cloud Model (LICM), the computing consumption of the proposed measurement algorithm is reduced by 50% on the premise of ensuring the stability and accuracy. It is proved to be feasible and effective by the application.

**Key words:** cloud model; similarity; overlap degree; logical relationship; measurement algorithm; time series

## 0 引言

20世纪90年代中期,李德毅院士在传统模糊数学和概率统计理论上提出了一种定量和定性相结合的数学模型——云模型,它把自然语言中定性概念的模糊性和随机性合理地结合在一起,实现了定性语言值与定量数值之间的相互映射。经过近二十年的研究发展,云模型已经扩展为云理论,其应用范围也涵盖了人工智能、数据挖掘、模糊控制和评价决策等领域。

在云理论中,不同的定性概念可用不同参数的云表示,但表达同类概念不同语言值的云却有多,如何来考察这些云之间的关联性,就涉及到云模型之间的相似程度<sup>[1]</sup>。文献[1]首次提出“相似云”的概念,云模型相似程度越大,则表示的定性概念一致性越高;反之,则一致性越低。

笔者通过查阅大量资料发现,传统云模型相似性判断方法有基于云滴数、向量相似和积分面积法等进行度量。例如,文献[1]是基于一定数量云滴间的距离来度量,然而选取云滴及对云滴的组合会导致时间复杂度过高。虽然,文献[2]对其进行了改进,但对云滴和阈值的依赖性导致计算结果不稳定的问题依然存在。文献[3]也面临同样的问题,这种不稳定性使得在相似性度量应用中受到制约。文献[4]用期望曲线或最大边界曲线所包围的公共面积作为度量相似性的标准,此方法解决了结果不稳定的问题,但在计算公共面积时复杂度还是较高。文献[5]把云模型数字特征作为一组向量,仅用向量相似性度量云模型相似性方法着实忽略了云模型独有的位置、形状和不确定性。

针对以上方法的不足,本文提出了一种基于云模型重叠度的相似性度量方法,在一定程度上克服了传统方法的不足。

收稿日期:2015-01-20;修回日期:2015-03-10。

基金项目:国家自然科学基金资助项目(61450011);山西省自然科学基金资助项目(2011011012-2)。

作者简介:孙妮妮(1988-),女,山东淄博人,硕士研究生,主要研究方向:智能信息处理、云模型; 陈泽华(1974-),女,山西神池人,副教授,博士,CCF会员,主要研究方向:智能信息处理、粒计算; 牛昱光(1958-),男,山西忻州人,副教授,主要研究方向:智能仪表、集散控制系统; 阎高伟(1970-),男,山西洪洞人,教授,博士,CCF会员,主要研究方向:智能信息处理、多传感器信息融合。

实验结果表明,这种方法能客观地度量云模型相似性、降低时间复杂度,并能提高数据分类效率及分类结果精确度。

## 1 云模型及其数字特征

云模型把概念的模糊性和随机性集成在一起,研究自然语言中最基本的语言原子所蕴含的不确定性的普遍规律,并从语言值表达的定性信息中获取定量数据的范围和分布规律,把精确值转换为恰当的定性语言值。

**定义1** 设  $U$  是一个用精确数值表示的定量论域,  $C$  是论域  $U$  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  $\mu(x) \in [0,1]$  是具有稳定倾向的随机数,则  $x$  在论域  $U$  上的分布称为云<sup>[6]</sup>,每个  $x$  称为一个云滴。

云模型用期望  $Ex$ 、熵  $En$  和超熵  $He$  三个数字特征反映所要表达概念的整体特性:期望(expected value)反映了定性概念的信息中心值,可认为是所有云滴在数域的重心位置;熵(entropy)反映了能够代表这一概念的云滴的离散程度,亦反映了论域空间中可被概念接受的云滴的取值范围;超熵(hyper entropy)是熵的不确定度量,反映了每个数值代表这个语言值确定度的凝聚性,也反映了云滴的凝聚性<sup>[7]</sup>。

由于正态分布的普适性,正态云是各种云模型中最重要的一种。而正态云的期望曲线是云理论研究数据集在空间中随机分布统计规律的重要方法,其一般方程为:

$$y = \exp[-(x - Ex)^2 / (2En^2)] \quad (1)$$

从图1可以看到,云模型的期望曲线光滑的穿过云滴“中间”,勾勒出云的整体轮廓,是云滴集合的“骨架”,所有的云滴都在期望曲线附近随机波动<sup>[8]</sup>,因此期望曲线可以很好地反映云的几何特征。

当超熵  $He$  较大时,云滴所呈现的特征明显区别于正态分布,云的期望曲线也不再明显。然而,由正态云模型雾化性质<sup>[9]</sup>可知,当  $He < En/3$  时,有 99.7% 的云滴落在最大边界曲线  $y_1 = \exp[-(x - Ex)^2 / (2(En + 3He)^2)]$  和最小边界曲线  $y_2 = \exp[-(x - Ex)^2 / (2(En - 3He)^2)]$  所围区域内<sup>[10]</sup>,此时云模型呈现良好的泛正态状态。故将  $En/3$  作为正态云模型的雾化点。

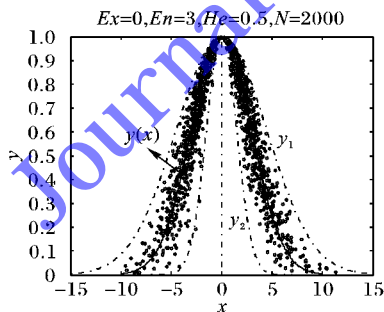


图1 正态云期望曲线及边界曲线

## 2 基于云模型重叠度的相似性度量

### 2.1 正态云模型的 $3En$ 规则

**定义2** 基础变量  $x$  中的任一小区间上的元素  $\Delta x$  对定性概念  $C$  的贡献  $\Delta D$  为:

$$\Delta D \approx \mu_C(x) \cdot \Delta x / \sqrt{2\pi}En \quad (2)$$

显然,论域上所有元素对概念  $C$  的总贡献  $D$  为:

$$D = \frac{\int_{-\infty}^{+\infty} \mu_C(x) dx}{\sqrt{2\pi}En} = \frac{\int_{-\infty}^{+\infty} \exp[-(x - Ex)^2 / (2En^2)] dx}{\sqrt{2\pi}En} = 1 \quad (3)$$

因为  $D = \frac{1}{\sqrt{2\pi}En} \int_{Ex-3En}^{Ex+3En} \mu_C(x) dx = 99.74\%$ , 所以对于论

域中的定性概念  $C$  的贡献的定量值,基本落在区间  $[Ex - 3En, Ex + 3En]$ ,甚至可以忽略区间之外的定量值对定性概念  $C$  的贡献,这就是正态云的  $3En$  规则<sup>[11]</sup>。

由式(1)及高斯函数曲线知识可得,云模型期望曲线也有  $3En$  规则。若将确定值  $\alpha = 0.0111$  记为  $3En$  规则处相应的确定度值,  $a$  和  $b$  分别为  $\alpha$  与期望曲线相交点(如图2所示),那么,相交所得左端点  $a$  即期望曲线  $y = \mu(x)$  定义域的下确界  $\inf\{C_\alpha\}$ ,右端点  $b$  为上确界  $\sup\{C_\alpha\}$ 。由此,可以给两个云作横向排序,来确定它们的位置关系和逻辑关系。

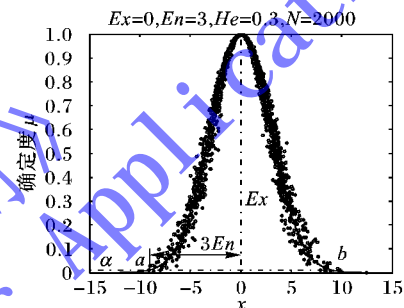


图2 云模型  $3En$  规则

**定义3** 云  $C_1$  和  $C_2$  是云模型论域空间  $U$  上的两个云,若存在:

$$\inf\{C_{1\alpha}\} < \inf\{C_{2\alpha}\} \text{ 且 } \sup\{C_{1\alpha}\} < \sup\{C_{2\alpha}\} \quad (4)$$

那么称云  $C_1$  小于云  $C_2$ ,记作  $C_1 < C_2$ 。

若存在:

$$\inf\{C_{1\alpha}\} \leq \inf\{C_{2\alpha}\} \text{ 且 } \sup\{C_{2\alpha}\} \leq \sup\{C_{1\alpha}\} \quad (5)$$

那么称云  $C_1$  包含云  $C_2$ ,记作  $C_1 \supseteq C_2$ 。

其中,  $\inf\{C_{i\alpha}\}$  和  $\sup\{C_{i\alpha}\}$  分别是云  $C_i$  期望曲线定义域的下确界和上确界(如图3所示)。

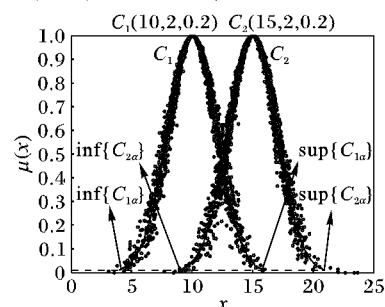


图3 云模型期望曲线定义域的上、下确界

若两个云是小于或包含关系,那么在云模型期望曲线  $3En$  规则下,两个云在横向上必然有重叠部分。为了定量描述两个云的重叠程度,引入了“云模型重叠度”的概念。

**定义4** 重叠度。假定  $C_1$  和  $C_2$  是云模型论域空间  $U$  上的两个云,若  $C_1 < C_2$ ,那么云  $C_1$  和  $C_2$  的重叠度定义为:

$$ol(C_{1\alpha}, C_{2\alpha}) = \frac{2 \cdot (\sup\{C_{1\alpha}\} - \inf\{C_{2\alpha}\})}{(\sup\{C_{1\alpha}\} - \inf\{C_{1\alpha}\}) + (\sup\{C_{2\alpha}\} - \inf\{C_{2\alpha}\})} \quad (6)$$

若  $C_1 \supset C_2$ ,那么云  $C_1$  和  $C_2$  的重叠度定义为:

$$ol(C_{1\alpha}, C_{2\alpha}) = \frac{2 \cdot (\sup\{C_{2\alpha}\} - \inf\{C_{2\alpha}\})}{(\sup\{C_{1\alpha}\} - \inf\{C_{1\alpha}\}) + (\sup\{C_{2\alpha}\} - \inf\{C_{2\alpha}\})} \quad (7)$$

## 2.2 基于期望曲线重叠度的相似云度量

目前的云模型相似性度量方法超高的时间复杂度,限制了云模型相似性度量的快速完成。可以借助云模型重叠度来刻画云模型的相似性。

根据文献[5]提供的方法计算云  $C_1(Ex_1, En_1, He_1)$  和云  $C_2(Ex_2, En_2, He_2)$  的交点  $x_1$  和  $x_2$ :

$$\begin{cases} x_1 = \frac{Ex_2 En_1 - Ex_1 En_2}{En_1 - En_2} \\ x_2 = \frac{Ex_1 En_2 - Ex_2 En_1}{En_1 + En_2} \end{cases} \quad (8)$$

根据交点  $x_1$  和  $x_2$  位置关系度量云模型相似性时,分以下3种情形。

情形1 若交点  $x_1$  和  $x_2$  全都位于  $3En$  规则区间之外,则云  $C_1$  和云  $C_2$  重叠度  $ol(C_1, C_2) = 0$ 。自然地,云模型相似度值  $Sim(C_1, C_2) = 0$ 。

情形2 若交点  $x_1$  和  $x_2$  中有一个位于  $3En$  规则区间之内,则云  $C_1$  和云  $C_2$  为小于关系,此时的重叠度  $ol$  由式(6)求得,云模型相似度值计算公式为:

$$Sim(C_1, C_2) = \frac{\mu - \alpha}{1 - \alpha} \cdot ol \quad (9)$$

其中  $\mu$  表示两条期望曲线交点对应的确定度值。

情形3 若交点  $x_1$  和  $x_2$  全都位于  $3En$  规则区间之内,则云  $C_1$  和云  $C_2$  是包含关系,由式(7)求得重叠度  $ol$ ,云模型相似度值计算公式为:

$$Sim(C_1, C_2) = \frac{\mu_{\max} - \alpha}{1 - \alpha} \cdot ol \quad (10)$$

其中  $\mu_{\max}$  为两交点对应确定度的较大值。

以上分析研究得到基于云模型期望曲线重叠度(Overlap based Expectation curve of Cloud Model, OECCM)的相似云度量算法。算法如下所示。

输入:两个云  $C_1(Ex_1, En_1, He_1)$ 、 $C_2(Ex_2, En_2, He_2)$ 。

输出:云模型相似度值  $Sim(C_1, C_2)$ 。

步骤1 先由式(8)求出两个云的交点  $x_1$  和  $x_2$ ,然后通过期望曲线解出相应的确定度值,比较大小后赋值给  $\mu_{\max}$  和  $\mu_{\min}$ 。

步骤2 若交点  $x_1$  和  $x_2$  都落在  $3En$  规则区间之外,则依据情形1得到云模型相似度值  $Sim(C_1, C_2) = 0$  且程序停止;否则,执行下一步。

步骤3 若交点  $x_1$  和  $x_2$  中有一个落在  $3En$  规则区间之内,两个云是小于关系则依据情形2以及式(9)求解重叠度  $ol$ ,得到云模型相似度值  $Sim(C_1, C_2)$  并输出;否则,执行下一步。

步骤4 若交点  $x_1$  和  $x_2$  同时落入  $3En$  规则区间之内,两个云是包含关系则依据情形3以及式(10)求解云模型重叠度  $ol$ ,得到云模型相似度值  $Sim(C_1, C_2)$  并输出。

## 2.3 基于最大边界曲线重叠度的相似云度量

云模型期望曲线能很好地反映正态云的整体特性,并从几何特征的角度来研究云模型相似性,进而忽略了超熵  $He$  的影响。实际情形中有时需要从局部角度来研究云模型相似性<sup>[4]</sup>,基于此提出了一种基于云模型最大边界曲线重叠度

(Overlap based Maximum boundary of Cloud Model, OMCM)的相似云度量算法。

随着超熵  $He$  的增大,云滴开始分散,正态云的期望曲线也不再明显。然而,却有 99.7% 的云滴落在云模型最大边界曲线  $y_1 = \exp[-(x - Ex)^2 / (2(En + 3He)^2)]$  和最小边界曲线  $y_2 = \exp[-(x - Ex)^2 / (2(En - 3He)^2)]$  所围的区域内。为避免减法可能出现“0”或者负数的情况,可选择最大边界曲线,研究云模型局部特征,取

$$y = \exp[-(x - Ex)^2 / (2(En + 3He)^2)] \quad (11)$$

对比式(1)和(11),令  $E = En + 3He$  便可得到与式(1)相似的式子,即:

$$y = \exp[-(x - Ex)^2 / (2E^2)] \quad (12)$$

接下来可按照 OECCM 算法思路得到基于最大边界曲线重叠度(OMCM)的相似云度量方法,完成云模型局部的相似性度量。

## 3 实验及分析

### 3.1 仿真实验

文献[1]设置的3个云的参数  $C_1(3, 3.123, 2.05)$ 、 $C_2(2, 3, 1)$  和  $C_3(1.585, 3.556, 1.358)$  得到云模型相似性的研究者的广泛应用和分析。故本文通过对这3个云模型进行数值仿真实验,并与文献[2]的基于区间的云相似度比较(Interval-Based Cloud Similarity Comparison, IBCSC)方法、文献[4]提出的基于期望曲线的云模型相似度(Expectation based Cloud Model, ECM)和基于最大边界曲线的正态云相似度(Maximum boundary based Cloud Model, MCM)计算方法以及文献[5]的基于云模型的相似度计算方法(Likeness comparing method based on Cloud Model, LICM)进行分析比较。通过仿真实验结果证明 OECCM 和 OMCM 方法的可行性和有效性,其实验结果如表1(由于云模型相似度的对称性  $Sim(C_1, C_2) = Sim(C_2, C_1)$ , 故为下矩阵形式)所示。

表1 各种相似性度量方法对3个云相似度的检验结果比较

方法	云参数	Sim/%		
		$C_1$	$C_2$	$C_3$
IBCSC 方法	$C_1$	97.02		
	$C_2$	95.59	97.69	
	$C_3$	95.19	97.55	98.41
ECM 方法	$C_1$	100.00		
	$C_2$	87.03	100.00	
	$C_3$	83.22	91.09	100.00
MCM 方法	$C_1$	100.00		
	$C_2$	78.49	100.00	
	$C_3$	89.55	88.02	100.00
LICM 方法	$C_1$	100.00		
	$C_2$	97.17	100.00	
	$C_3$	94.38	98.50	100.00
OECCM 方法	$C_1$	100.00		
	$C_2$	90.60	100.00	
	$C_3$	87.40	91.33	100.00
OMCM 方法	$C_1$	100.00		
	$C_2$	78.40	100.00	
	$C_3$	89.96	88.00	100.00

由表1可以发现,基于最大边界曲线重叠度的 OMCM 方



法与 MCM 方法结果一致。然而,相似性结果与其他方法有些许差距,其主要原因是:当超熵  $He > En/3$  时,部分云滴脱离了最大和最小边界曲线所夹范围<sup>[4]</sup>。所以,当云模型呈现泛正态时,OMCM 方法的适用范围会大大增加。

除此之外,OECM 方法得出的结论与 IBCSC、LICM、ECM 等传统方法一致。OECM 方法与 IBCSC 方法相比,不仅结果稳定性强,而且得到的相似性结果容易区分;与 LICM 方法相比,将期望  $Ex$ 、熵  $En$  和超熵  $He$  对云模型相似性度量的贡献分别作考虑,综合了云模型独有的位置和形状差异;与 ECM 和 MCM 方法相比,解决了由于曲线积分计算而导致的时间复杂度过高的问题。

### 3.2 时间序列分类实验

时间序列数据有严格的时间先后顺序,是时间序列数据分析和数据挖掘中的重要任务之一。此次时间序列分类实验使用 UCI 中的时间序列数据集 (synthetic control chart dataset)<sup>[12]</sup>,此数据集是 1999 年在 Alcock 和 Manolopoulos 产生的 600 例综合控制图,分为 6 类数据,每组数据都有 100 个长度为 60 的时间序列<sup>[13]</sup>。此次实验,从分析算法分类正确率和时间复杂度入手,对比了各个云模型相似性度量方法在时间序列分类实验的计算结果。

研究各算法分类正确率时,基于最近邻分类 ( $K$ -Nearest Neighbors, KNN) 算法<sup>[14]</sup>和协同过滤推荐系统<sup>[5]</sup>思想,对分类标准逐步提高:事先规定最接近点数  $K=10$ ,然后根据相似性算法结果所属类别个数从 1 到 10 依次计算,得到分类结果如图 4 所示。分析时间复杂度时,从总数据中任意抽取数据个数分别为 60, 180, 300, 420, 540 和 600 的数据集。6 种情况下,5 种相似性方法 (IBCSC 方法和最大最小贴近度 (Maximum and Minimum based Cloud Model, MMCM) 方法<sup>[3]</sup>是基于云滴数的计算方法,时间复杂度远高于其他算法,故没有在图中标注) CPU 所耗时间结果如图 5 所示。

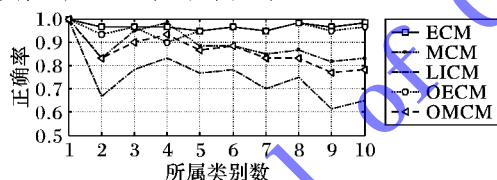


图4 5种相似性方法正确率对比

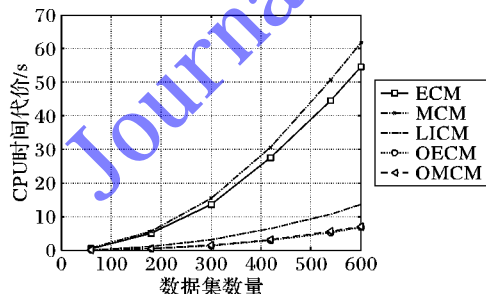


图5 5种相似性方法的 CPU 时间消耗对比

图 4 表明:OECM 与 ECM 方法对时间序列分类的正确率不分上下,但 OECM 方法的稳定性略强;OMCM 和 MCM 方法分类结果稍稍逊色于 OECM 和 MCM 方法;LICM 方法对时间序列的分类结果再次之,欠稳定。然而,图 5 显示的 CPU 时间消耗中:OECM 和 OMCM 方法时间优势明显 (ECM 和 MCM 是基于曲线积分求公共面积的);其中,OECM 和 OMCM 方法的时间复杂度比 ECM 降低了 88%,与目前计算消耗较低的

LICM 方法相比,时间复杂度也降低了 50%。结合时间序列分类正确率和 CPU 时间代价,验证了 OECM 和 OMCM 方法的可行性和有效性。

## 4 结语

云模型相似性度量是云模型在系统推荐和评估领域亟待深入研究和解决的问题,而相似云的提出是对云模型理论的丰富和扩展。本文从云模型的横向基础上,明确了两个云模型的小于关系和包含关系,综合云模型的位置和逻辑关系定义了重叠度,最终确定了一种度量云模型相似性的新方法。对文献[1]中的 3 个云模型进行相似度仿真实验,并将算法应用到时间序列分类实例中,得到的实验结果不仅稳定且时间复杂度低,说明该算法在推荐系统和大数据处理方面有一定的应用前景。

将本文算法应用到其他领域是下一步的研究方向,同时,OMCM 方法对超熵的依赖性也是需要优化的问题。

### 参考文献:

- [1] ZHANG Y, ZHAO D, LI D. The similar cloud and the measurement method [J]. Information and Control, 2004, 33(2): 130 - 132. (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.)
- [2] CAI S, FANG W, ZHAO J, et al. Research of interval-based cloud similarity comparison algorithm [J]. Journal of Chinese Computer Systems, 2011, 32(12): 2457 - 2460. (蔡绍滨, 方伟, 赵靖, 等. 基于区间的云相似度比较算法的研究[J]. 小型微型计算机系统, 2011, 32(12): 2457 - 2460.)
- [3] JIN L, TAN S. Similarity measurement between cloud models based on close degree [J]. Application Research of Computers, 2014, 31(5): 1309 - 1311. (金璐, 覃思义. 基于云模型间贴近度的相似度量法[J]. 计算机应用研究, 2014, 31(5): 1309 - 1311.)
- [4] LI H, GUO C, QIU W. Similarity measurement between normal cloud models [J]. Acta Electronica Sinica, 2011, 39(11): 2562 - 2567. (李海林, 郭崇慧, 邱望仁. 正态云模型相似度计算方法[J]. 电子学报, 2011, 39(11): 2562 - 2567.)
- [5] ZHANG G, LI D, LI P, et al. A collaborative filtering recommendation algorithm based on cloud model [J]. Journal of Software, 2007, 18(10): 2404 - 2411. (张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推荐算法[J]. 软件学报, 2007, 18(10): 2404 - 2411.)
- [6] LI D, MENG H, SHI X. Cloud and cloud generator [J]. Computer Research and Development, 1995, 32(6): 15 - 20. (李德毅, 孟海军, 史雪梅. 隶属云和隶属云发生器[J]. 计算机研究与发展, 1995, 32(6): 15 - 20.)
- [7] LI D. Uncertainty in knowledge representation [J]. Engineering Science, 2000, 2(10): 74 - 79. (李德毅. 知识表示中的不确定性[J]. 中国工程科学, 2000, 2(10): 74 - 79.)
- [8] MIAO D, WANG G, YAO Y, et al. Cloud model and granular computing [M]. Beijing: Science Press, 2012: 4 - 30. (苗夺谦, 王国胤, 姚一豫, 等. 云模型与粒计算[M]. 北京: 科学出版社, 2012: 4 - 30.)
- [9] LIU Y, LI D. Statistics on atomized feature of normal cloud model [J]. Journal of Beijing University of Aeronautics and Astronautics, 2010, 36(11): 1321 - 1324. (刘禹, 李德毅. 正态云模型雾化性质统计分析[J]. 北京航空航天大学学报, 2010, 36(11): 1321 - 1324.)

(下转第 1964 页)

样本既具有一定的随机性,又保持了原始样本的概率分布,从而有效提升了数据质量,解决了数据不平衡现象严重影响小类预测精度的问题。

2)融合用户信息、社交关系和主题信息的集成学习方法,明显提高了微博转发行为预测的精度,证明本文方法在解决面向主题的微博行为预测问题时是有效而且可行的。

#### 参考文献:

- [1] WANG Y, JIN X, CHENG X. Network big data: present and future [J]. Chinese Journal of Computers, 2013, 36(6): 1125 - 1138. (王元卓,靳小龙,程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125 - 1138.)
- [2] HUANG Y, SUN X, LIU Z, *et al.* The microblog retweeting prediction evaluation system and performance comparison [J]. Journal of Harbin University of Science and Technology, 2013, 18(4): 52 - 57. (黄英来,孙晓芳,刘镇波,等. 微博转发预测算法评测系统的建立及性能比较[J]. 哈尔滨理工大学学报, 2013, 18(4): 52 - 57.)
- [3] SUH B, HONG L C, PIROLI P, *et al.* Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network [C]// Proceedings of the 2010 IEEE International Conference on Social Computing. Piscataway: IEEE, 2010: 177 - 184.
- [4] XU Z, YANG Q. Analyzing user retweet behavior on twitter [C]// Proceedings of 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Piscataway: IEEE, 2012: 46 - 50.
- [5] ROMERO D M, MEEDER B, KLEINBERG J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter [C]// Proceedings of the 20th International Conference on World Wide Web. New York: ACM, 2011: 695 - 740.
- [6] WENG J, LIM E P, JIANG J. TwitterRank: finding topic-sensitive influential twitterers [C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010: 261 - 270.
- [7] WELCH M J, SCHONFELD U, HE D. Topical semantics of twitter links [C]// Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011: 327 - 336.
- [8] MORCHID M, DUFOUR R, LINARES G, *et al.* Feature selection using principal component analysis for massive retweet detection [J]. Pattern Recognition Letters, 2014, 49(11): 33 - 39.
- [9] PENG H, ZHU J, PIAO D Z, *et al.* Retweet modeling using conditional random fields [C]// ICDMW'11: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. Washington, DC: IEEE Computer Society, 2011: 336 - 343.
- [10] ZHANG Y, LU R, YANG Q. Predicting retweeting in microblogs [J]. Journal of Chinese Information Processing, 2012, 26(4): 109 - 114. (张畅,路荣,杨青. 微博客中转发行为的预测研究[J]. 中文信息学报, 2012, 26(4): 109 - 114.)
- [11] LI Y, YU H, LIU L. Predict algorithm of micro-blog retweet scale based on SVM [J]. Application Research of Computers, 2013, 30(9): 2594 - 2597. (李英乐,于洪涛,刘力雄. 基于SVM的微博转发规模预测方法[J]. 计算机应用研究, 2013, 30(9): 2594 - 2597.)
- [12] XIE J, LIU G, SU B, *et al.* Prediction of user's retweet behavior in social network [J]. Journal of Shanghai Jiaotong University, 2013, 47(4): 584 - 588. (谢婧,刘功申,苏波,等. 社交网络中的用户转发行为预测[J]. 上海交通大学学报, 2013, 47(4): 584 - 588.)
- [13] LUO Z, WU X, CAI W, *et al.* Examining multi-factor interactions in microblogging based on log-linear modeling [C]// Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York: ACM, 2012: 189 - 193.
- [14] LUO Z, CAI W, CHEN T. Microblogging retweet prediction algorithm based on random forest [J]. Computer Science, 2014, 41(4): 62 - 64. (罗知林,蔡皖东,陈挺. 一种基于随机森林的微博转发预测算法[J]. 计算机科学, 2014, 41(4): 62 - 64.)
- [15] FANG X, ZHANG H, GAO S. Web spam detection based on SMOTE and random forests [J]. Journal of Shandong University: Engineering Science, 2013, 43(1): 22 - 26. (房晓南,张化祥,高爽. 基于SMOTE和随机森林的Web spam检测[J]. 山东大学学报: 工学版, 2013, 43(1): 22 - 26.)
- [16] YU H, GAO S, ZHAO J, *et al.* Classification for imbalanced microarray data based on oversampling technology and random forest [J]. Computer Science, 2012, 39(5): 190 - 194. (于化龙,高尚,赵靖,等. 基于过采样技术和随机森林的不平衡微阵列数据分类方法研究[J]. 计算机科学, 2012, 39(5): 190 - 194.)
- [17] LIAN J, ZHOU X, CAO W, *et al.* SINA microblog data retrieval [J]. Journal of Tsinghua University, 2011, 51(10): 1300 - 1305. (廉捷,周欣,曹伟,等. 新浪微博数据挖掘方案[J]. 清华大学学报: 自然科学版, 2011, 51(10): 1300 - 1305.)
- [18] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5 - 32.
- [19] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2): 123 - 140.

(上接第1958页)

- [10] LIU C, LI D, DU Y, *et al.* Statistics on atomized feature of normal cloud model [J]. Journal of Beijing University of Aeronautics and Astronautics, 2005, 34(2): 236 - 239. (刘常昱,李德毅,杜鹤,等. 正态云模型的统计分析[J]. 信息与控制, 2005, 34(2): 236 - 239.)
- [11] LI D, DU Y. Artificial intelligence with uncertainty [M]. Beijing: National Defense Industry Press, 2005: 237 - 248. (李德毅,杜鹤. 不确定性人工智能[M]. 北京: 国防工业出版社, 2005: 237 - 248.)
- [12] PHAM D, CHAN A. Control chart pattern recognition using a new type of self organizing neural network [J]. Journal of Systems and Control Engineering, 1988, 212(2): 115 - 127.
- [13] KANNAN S R, RAMATHILAGAM S, CHUNG P C. Effective fuzzy c-means clustering algorithms for data clustering problems [J]. Expert Systems with Applications, 2012, 39(7): 6292 - 6300.
- [14] BELHAOUARI S. Fast and accuracy control chart pattern recognition using a new cluster-k-nearest neighbor [J]. International Scholarly and Scientific Research & Innovation, 2009, 3(1): 970 - 974.