决策连续形式背景的可视化数据离散化方法*

张 涛，师浩斌，李 林，李朝辉
（燕山大学 信息科学与工程学院，河北 秦皇岛 066004）

摘 要： 连续形式背景离散化是形式概念分析领域重要的基础问题之一。针对形式背景离散化的特殊要求，提出了一种可视化的数据离散化方法。该方法借助可视化方法对数据类别分布进行表示，将连续数据分布转换为图形分布，进一步利用视觉模糊性对图形空间进行处理，进而将决策连续背景离散化。通过 UCI 数据集上的实验表明，与传统离散化方法相比，采用该方法进行数据离散化后的二值形式背景具有结构简单且不失准确性的优点。

关键词： 形式背景；离散化；可视化；形式概念分析
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2016)02-0388-04
doi:10.3969/j.issn.1001-3695.2016.02.015

Visual discretization for decision continuous formal context

Zhang Tao, Shi Haobin, Li Lin, Li Zhaohui
(*School of Information Science & Engineering, Yanshan University, Qinhuangdao Hebei* 066004, *China*)

Abstract: Discretization for decision continuous formal context is one of basic issues in the field of formal concept analysis. According to the special requirements of formal concept analysis field for data discretization, this paper presented a method which is in the form of visualization. This method presented the data distribution by visualization, converted the data distribution to figure distribution, and fuzzy analysis the figure to discretize the decision continuous formal context. The whole progress considered not only the category distribution, but also the fuzzy of the data. The experiments show that compared to the traditional discretization methods, the binary formal context after using this method to discretize data has the advantages of simple structure and without loss of accuracy.

Key words: formal context; discretization; visualization; formal concept analysis

0 引言

形式概念分析(formal concept analysis,FCA)是一种研究数据间结构和逻辑的数学方法。该理论于 1982 年由德国的 Wille[1] 首先提出。由于其同时具有本体描述与逻辑推理能力，其目前已经广泛地应用于机器学习[2]、数据挖掘[3]、知识发现[4]、信息检索[5]、数据抽取[6]等诸多领域。

在形式概念分析理论中，形式背景(formal context)是其主要处理对象与数据基础[7]。目前对形式背景的分析理论主要以二值背景为主，多值背景则通过平凡运算转换为二值背景进行处理[7]。因此形式概念分析具有较强的定性数据分析能力。而在机器学习等应用领域中，其属性值大多数为定量数据，且为一个值域范围内的连续分布[8]。因此根据数据分布情况将连续值数据集进行离散化，使其转换为多值或二值形式背景成为了形式概念分析应用领域需要解决的基本问题[9]。

在传统的形式背景转换方法中，根据整个属性空间的整体性或局部性可以分为全局离散化和局部离散化；根据离散化计算时对类别信息参考与否分为有监督离散化和无监督离散化；根据离散化与决策生成的串并行关系分为静态离散化和动态离散化；根据对最小离散区间的单元定义与处理分为归并离散化和拆分离散化；根据离散化过程是否递进完成分为直接离散化和递进离散化[10]。

各种离散化方法均有其自身的特点。经典的离散化方法大多直接来自数据挖掘等传统领域，其对于数据挖掘算法具有较好的应用。但对于以二值背景为处理对象的形式概念分析而言，其存在着直观性不足、离散后信息缺失、过度离散造成概念格生成复杂难以应用等问题，这些因素大大制约了形式概念分析在连续数值领域的应用。

针对传统数据离散化方法的不足，本文提出可视化数据离散化方法，从可视化的角度完成满足形式结构表示要求的数据离散化过程。与一般的离散化过程不同，该方法借助可视化方法对数据类别分布进行表示，利用视觉模糊性对决策连续背景进行离散化。不但对数据的空间分布与类别分布进行统一考虑，同时兼顾量化等级的数量，以降低后期分析的复杂度，提高概念格等分析工具的生成效率。

1 形式背景与决策形式背景

形式背景是形式概念分析的数据基础[7]，一个形式背景

收稿日期：2014-09-27；修回日期：2014-11-10 基金项目：国家自然科学基金资助项目(61273019)；河北省自然科学基金资助项目(F2015203013)；河北省社会科学基金年度项目(HB14YY005)；国家社会科学基金资助项目(12BYY121)

作者简介：张涛(1979-)，男，河北唐山人，副教授，博士，主要研究方向为形式概念分析、可视化知识发现(zhtao@ysu.edu.cn)；师浩斌(1990-)，男，硕士研究生，主要研究方向为形式概念分析；李林(1977-)，男，副教授，博士，主要研究方向为数据库分析；李朝辉(1981-)，副教授，博士，主要研究方向为神经信息处理.

$K:=(G,M,I)$ 是由两个集合 G 和 M 以及 G 与 M 间的关系 I 组成。G 的元素称为对象,M 的元素称为属性。gIm 表示对象 g 具有属性 m。关系 I 也称为背景关联的关系,可以用 $g\not I m$ 表示 $(g,m)\notin I$。

在该形式背景定义基础上对数据进行推广,可得多值形式背景的概念[7]。多值背景是一个四元组 (U,M,W,I),其中 U 是对象的集合,M 是属性的集合,W 是属性值的集合,$I\subseteq U\times M\times W$ 是 U、M、W 上的关系。I 应该满足 $(u,m,w_1)\in I$,$(u,w,m_2)\in I$,就必有 $w_1=w_2$。如果 $(u,m,w_1)\in I$,则表示对象 u 在属性 m 上取值为 w。

对于多值背景,可采用标尺进行二值化[8]。标尺是一类特殊的背景 (Us,Ms,Is),用这个标尺来解释 (U,M,W,I) 中的属性 m 的多值是指利用这个标尺产生一个新的背景 (U,Ms,J),其中 $(u,m)\in J$ 等价于 $\exists w\in W,(u,m,w)\in I,(c,w)\in I_s$。由于利用标尺完成多值背景与二值背景的转换在文献[11]中有详细描述,在此不再赘述。

无论是形式背景还是多值形式背景,其描述的都是对象与属性间的关系。而属性中的蕴涵关系则可表示为

定义 1[9]　设 $r=\{u_1,\cdots,u_n\}$ 是关系模式 $R=\{A_1,\cdots,A_m\}$ 上的一个关系。令 $M=\underset{A\in R}{U}\mathrm{dom}(A)$,这里 $\mathrm{dom}(A)$ 是属性 A 的值域,$U=r=\{u_1,\cdots,u_n\}$,$I=\{(u,m)\mid\exists A_i\in R,u[A_i]=m\}$。这里 $u\in U,m\in M$,则称 (U,M,I) 是关系 r 对应的形式背景。若 $Y_1,Y_2\subseteq M$,则称表达式 $Y_1\to Y_2$ 为一个值依赖,当 $g(Y_1)\subseteq g(Y_2)$ 时,称 $Y_1\to Y_2$ 在背景 (U,M,I) 中成立。

定义 2　设 (U,M,I) 是一个背景,$A\to B$ 是一个值依赖,$T\subseteq M$,则当且仅当 $A\not\subset T$ 或 $B\subseteq T$ 时称 T 与 $A\to B$ 相关,与 $A\to B$ 所有相关集合的族记做 $\varepsilon(A\to B)$。若 Σ 是值依赖集合,T 与 Σ 中的每一个值依赖都相关,则称 T 与 Σ 相关。与 Σ 所有相关集合的族记做 $\varepsilon(\Sigma)$。若 $T_1,T_2,\cdots\subseteq M$,而 $A\to B$ 与 $T_1,T_2,\cdots$ 都相关,则称 $A\to B$ 在 $\{T_1,T_2,\cdots\}$ 中成立。

显然,值依赖从属性关系角度定义了属性间的相互关系。决策形式背景是值依赖的重要类型[12]。称五元组 (U,M,I,D,J) 为一个决策形式背景,U 为对象集,M 为条件属性集,A 为区间函数,表示条件属性 M 的定义域,D 为决策属性集。显然,在形式背景 (U,A,I,D,J) 中,其子背景 (U,A,I) 和 (U,D,J) 也为形式背景。

五元组 (U,M,I,D,J) 表示的背景仍为二值背景。以此为基础,当条件属性为连续值时,其可采用六元组 (U,M,A,I,D,J) 进行表示,称为决策连续形式背景。A 为区间函数,表示条件属性 M 的定义域。

在数据集合中,设训练样本集合 $X=\{x_1,x_2,\cdots,x_m\}$,共有 m 个训练样本,每个样本可表示为 $x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}\in R^d$。训练样本类别标签集合为 $C=\{c_1,c_2,\cdots,c_k\}$,显然,$k\leqslant m$。设表示训练样本 x_i 的类别为 $L(x_i)$,则必有 $L(x_i)\in C$。由于标签集合 C 本身为离散值,所以其在形式背景 (U,M,A,I,D,J) 中,$D=C,J=L(x_i)$。决策形式背景的离散化即根据背景 (U,D,J) 对数据集 $X=\{x_1,x_2,\cdots,x_m\}$ 进行离散化,从而获得完整的五元组 (U,M,I,D,J) 的过程。从形式背景变换角度看,即将六元组 (U,M,A,I,D,J) 变为五元组 (U,M,I,D,J),使得形式概念计算可完成的数据转换过程。

2　可视化离散化

2.1　数据空间的色度学可视化

为了在不影响空间分布的情况下进行数据类别表示,本文采用文献[13,14]中的色度学可视化表示方法。

对于一个已知类别数据 $u_i=\{x_{i1},x_{i2},\cdots,x_{id}\}\in U,L(x_i)\in c_m$。其第 j 个特征在笛卡尔空间中映射空间坐标为 (j,x_{ij})。将色度信息作为一个维度进行统一表示,由于色度空间考虑了类别分布,该特征在空间中的色度值可表示为 $f(j,x_{ij},c_m)$。其中 c_m 维表示类别,在可视空间内以色度进行表示,不影响原有的空间结构与表示过程。

对于多个数据所组成的数据集,其特定坐标下的色度学表示可通过色度学合成完成。设在该坐标 (x_i,x_j) 下共有 l 个对象,需要 l 个基色进行表示。设选取的基色在颜色空间坐标为 $h_k(r_k,\theta_k)$,幅值 r_k 表示饱和度,用于表示类别的混合程度。对于基色,由于表示单一类别,所以饱和度最大,可令 $r_k=1$。相角 θ_k 表示色调,不同的色调对应不同相角。

在决策背景的离散过程中,类别的概率分布是决策的主要依据,而样本的绝对数目分布对于决策过程的影响包含在后期形式概念分析过程中。因此,本文对空间中相同坐标点的不同类别数据作归一化处理,即

$$f(x_i,x_j,c_k)=\frac{f(x_i,x_j,c_k)}{\sum_{k=1}^{l}f(x_i,x_j,c_k)}\tag{1}$$

在色度学合成中,相同空间上混合色的色调 θ 和色饱和度 r 分别为

$$\theta=\sum_{k=1}^{l}f(x_i,x_j,c_k)\,\theta_k\tag{2}$$

$$r=\max\left(\sum_{k=1}^{l}f(x_i,x_j,c_k)\,r_k,1\right)\tag{3}$$

通过色度学计算,将类别信息转换为色度信息对当前像素点进行着色。

2.2　可视化空间离散化

通过可视化表示,将数据表示成可视空间中的点分布。由于数据采集本身的离散特性及其采集样本数量特性,类别数据在可视化空间中将表现为线段分布。根据空间点或线的颜色特征对其进行离散化划分,在保证分类精度的同时满足区段最小化,是本方法的核心思想。

对于某一特征数据 $a=\{x_{1i},x_{2i},\cdots,x_{ni}\}$,为表示方便,其值域根据其数值大小非递减排列,则其值域 $V_a=\{v_a^1,v_a^2,\cdots,v_a^i,\cdots,v_a^n\}$,则有

$$v_a^1<\cdots<v_a^i<\cdots<v_a^{|v_a|}\tag{4}$$

则在形式背景 (U,A,I) 中,数值为 v_a^i 的特征的集合可表示为

$$X_a^i=\{x\in a\mid x=v_a^i\}\tag{5}$$

结合子背景 (U,D,J),集合 X_a^i 对应的类别可表示为

$$\Delta_a^i=\{d\in D\mid\exists x\in X_a^i,L(x)=d\}\tag{6}$$

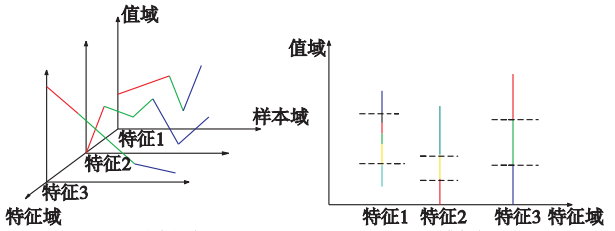
则值域相邻的不同类别间隔为

$$C_a=\left\{\frac{v_a^i+v_a^{i+1}}{2}\;\middle|\;|\Delta_a^i|>\mathrm{lor}\,|\Delta_a^{i+1}|>\mathrm{lor}\Delta_a^i\neq\Delta_a^{i+1}\right\}\tag{7}$$

在以划分为代表的离散化方法中,其对 v_a^i 和 v_a^{i+1} 取中间间隔,以此作为离散化的区间标准。从投影角度看,该离散化过程是将原有的值域—样本域—特征域三维空间向值域—特征域平面的投影过程,如图 1 所示。

从图 1 中可以直观观察到,该方法的本质是根据值域分布

与类别分布对数据进行最大化离散，从而保证每个量化区间内类别数目最小，满足后期分类性能的要求。但对于类别混叠严重的数据，过细的离散化不但使离散化后的属性过多，导致形式结构分析阶段计算复杂度增加，而且容易形成过学习，使分类性能下降。



(a) 原始分布 (b) 二维投影

图1 离散化的映射表示

为了解决划分过细的问题，可通过对值域—特征域平面的投影对数据进行聚合，其基本原则为：设定量化间隔最小阈值 d，若 $\Delta_a^i < d$，则该量化间隔与相邻间隔中较小的进行合并，从而将小的量化阶模糊化，形成基于类别分布的大阶段量化。依主动生长原理可知[15]，该量化后的区间间隔

$$C_a^i = \begin{cases} C_a^i & C_a^i > d \text{ 且 } C_a^{i+1} > d \\ C_a^i + C_a^{i+1} & C_a^i > d \text{ 且 } C_a^{i+1} < d \\ C_a^i + C_a^{i+1} & C_a^i < d \text{ 且 } C_a^{i+1} > d \\ C_a^i + C_a^{i+1} & C_a^i < d \text{ 且 } C_a^{i+1} < d \end{cases} \tag{8}$$

由分布条件可将式(8)合并为

$$C_a^i = \begin{cases} C_a^i & C_a^i > d \text{ 且 } C_a^{i+1} > d \\ C_a^i + C_a^{i+1} & \text{其他} \end{cases} \tag{9}$$

按主动生长原理进行简化[15]，式(9)可进一步简化为

$$C_a^i = C_a^i - \mathrm{Sign}(\mathrm{Sign}(\mathrm{Sign}(C_a^v - d) + \mathrm{Sign}(C_a^i - d)) - 1) \times C_a^{i+1} \tag{10}$$

以此为依据对原始投影空间进行区间融合，如图 2 所示。特征 1 数据的中段数据，由于每段间隔过小，所以被合并为一个模糊区间；特征 2 数据中，黄色和青色(见电子版)均表示类别混叠区域，但由于混叠类别不同且各段间隔较大，所以分别进行量化；特征 3 中均为单独类别区域且各段间隔较大，可以直接根据原始结果进行量化。

2.3 形式背景生成

利用以上方法对各特征进行可视化数据离散化，可将六元组(U,M,A,I,D,J)中的属性 A 由连续区间表示变为区间段表示，即另一个连续集合变为有限集合的过程，以此形成背景 $K=(U,M,W,I,D,J)$，其中 U、M、D、J 均与原表示相同。W 为属性值域，此处为离散化后的特征值；$I \subseteq U \times M \times W$ 为三元关系序偶，满足当$(u,m,w) \in R$ 且$(u,m,v) \in R$ 时有 $w=v$ 成立。以此将连续形式背景转换为多值形式背景。

进一步，可利用标尺法或平凡运算将多值背景转换为二值背景[7]，最终获得将六元组的决策连续形式背景(U,M,A,I,D,J)变为决策二值形式背景(U,M,A,I,D,J)的过程。

3 实验

3.1 概念格生成实验

为了验证本文方法的有效性，本文采用机器学习中常用的 Iris 数据为例进行划分。按第 2 章所述方法，其划分间隔如图 3 所示。

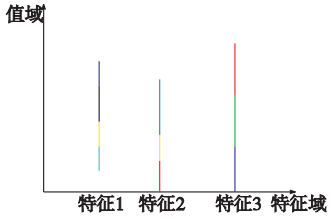


图2 不同混叠数据的离散化划分示意图

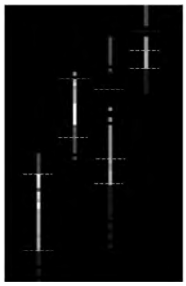


图3 Iris数据集特征的离散化区间

通过对图 3 的分析可知，利用本文方法，可以将 Iris 数据进行可视化表示。从可视化图形中，可以直观发现各特征下类别的混叠程度。通过对图形空间的划分，完成对单类数据与混叠数据的模糊划分，进而完成连续背景向多值背景的转换。获取量化间隔结果如表 1 所示。

表 1 Iris 特征量化间隔

原始特征	量化后特征	原始特征	量化后特征
sepal length	A [*，4.9) A [4.9，7.1) A [7.1，*)	petal length	C [*，2.5) C [2.5，4.5) C [4.5，5.2) C [5.2，*)
sepal width	B [*，2.1) B [2.1，3.9) B [3.9，*)	petal width	D [*，0.8) D [0.8，1.4) D [1.4，1.9) D [1.9，*]

设离散后形成的形式背景为 $K=(O,A,W,R)$，在该背景中，O 与原始 Iris 数据集中的对象集合相同，共有 150 个对象；$A=\{\text{sepal length、sepal width、petal length、petal width}\}$；$W$ 为离散化后的特征值，对应于表 1 中离散化后的特征；$R \subseteq O \times A \times W$ 为三元关系序偶。将该形式背景进行平凡运算，形成 14 个二值属性的形式背景。如表 2 所示。

表 2 离散化后的形式背景

标号	对象类别	A [*，4.9)	A [4.9，7.1)	A [7.1，*)	B [*，2.1)	B [2.1，3.9)	B [3.9，*)	C [*，2.5)	C [2.5，4.5)	C [4.5，5.2)	C [5.2，*)	D [*，0.8)	D [0.8，1.4)	D [1.4，1.9)	D [1.9，*)
1	setosa		x			x		x				x			
2	setosa	x				x		x				x			
3	setosa		x				x	x				x			
4	versicolor		x			x				x				x	
5	versicolor		x			x			x				x		
6	versicolor		x			x				x			x		
7	versicolor		x			x			x					x	
8	versicolor		x		x				x				x		
9	virginica		x			x				x				x	
10	virginica		x			x					x				x
11	virginica		x			x				x					x
12	virginica			x		x					x				x
13	virginica		x			x					x			x	
14	virginica			x		x					x			x	

该背景的概念格如图 4 所示。

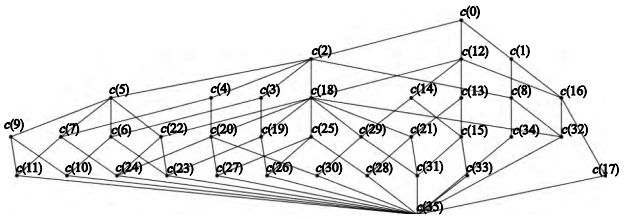


图4　原始概念格表示

本实验证明了本文所提方法在决策联系形式背景离散化过程中的可行性。以下通过对比实验验证本文方法与经典方法相比的优缺点。

3.2　粒度划分实验

为了验证本文算法的有效性,利用 UCI 数据库中的多个典型数据集进行对比测试。所选择的数据集及其属性如表 3 所示。其中,Iris 与 wine 数据集为模式分类的常用测试集,分别测试低维特征与高维特征情况下的多类分类器的分类性能,而 glass 数据集中,特征维数相对较高,且类别数目较高,可以表现多类分类性能。实验中用到的这些数据集来自物理科学与生命科学领域,具体的应用包括了产品分类和物种识别,且均为实际测量的实验数据,含有一定的测量误差,因此可以在一定程度上代表分类器在实际应用中的分类性能。

表 3　实验用到的各数据集属性

数据集	所属领域	条件属性数	决策属性数目	样本数
Iris	生命科学	4	3	150
wine	物理科学	13	3	178
glass	物理科学	9	6	214

在对比算法中,分别采用自然离散化、布尔推理、信息熵和等频离散化作为对比算法。其中,自然算法根据需要离散化的某属性值将对象排序,按照对象的顺序,只要对象的决策值改变,就产生一个新区间,该算法产生保持决策系统一致性水平所需的所有分割点。布尔推理算法[16]是根据粗糙集理论和布尔推理提出的一种离散化方法,并对其进行了一定的改进,这是一种全局有监督算法。利用布尔推理程序,可以合并一些初始分割点,剩下的分割点是保留决策系统不可分辨力的最小集合。信息熵利用类别信息继续区间划分。等频率离散化则是将待离散化属性的值域范围划分成若干个子区间,使每个子区间包括大致相等的数目。

通过表 4 可知,与经典离散化方法相比,本文方法在属性数目上与等频离散方法基本相当。其原因在于离散过程中通过色度空间计算与模糊控制,部分实现了频度计算的过程,因此在属性数目上相当。其明显优于自然离散余信息熵方法。其原因在于自然离散与信息熵在离散过程中不对离散后的属性数目作约束,但数据类别分布混叠严重时其属性数目将急剧增大。与布尔推理相比,本文属性数目较为稳定,因此具有良好的泛化能力。

表 4　不同离散化方法后二值条件属性对比

类别	自然离散	布尔推理	信息熵	等频离散	本文方法
Iris	60	11	35	12	15
wine	723	722	150	39	42
glass	690	27	263	27	35

由于本文离散化方法针对概念格设计,所以其测试也在概念格中进行。将测试样本的条件属性与概念格中的概念进行比对,获得测试样本的类别。在实验过程中,为了确保分类性能更为客观,并避免训练集和测试集的依赖,分类器精度的估计采用留一法交叉验证(leave one out cross validation, LOOCV)。留一法是指设数据集共有 N 个样本,使用($N-1$)个样本设计分类器,并估计剩余的一个样本; 对于训练集重复 N 次。这种估计虽然计算量大,但是无偏的。根据以上的实验数据与实验条件,得到的结果如表 5 所示。

表 5　不同离散化方法分类精度对比　/%

类别	自然离散	布尔推理	信息熵	等频离散	本文方法
Iris	96.67	88.67	95.32	98.33	95.32
wine	89.89	84.27	92.69	87.08	90.12
glass	77.10	65.42	66.35	71.03	75.28

通过表 5 可知,本文方法与自然离散基本相当,但从表 4 可知,自然离散方法由于属性数目多,其计算复杂度偏高。本文方法形成的概念格在分类精度上优于布尔推理和等频离散方法,其原因在于其在离散化过程中,不但考虑了数据的空间分布,同时考虑了类别分布。另外,通过可视化表示,引入了模糊处理思想,使得其本身已经具有一定的分类能力; 信息熵方法虽然在离散过程中考虑到了类别分布问题,但其并不适用于概念格构造下的分类器,因此分类精度并未体现优势。

4　结束语

本文提出了一种基于可视化方法的决策连续形式背景的数据离散化方法。该方法通过对数据空间、类别的综合分析获得离散间隔。在保证分类精度的同时针对概念格设计了离散间隔方法,为连续数据集下的形式概念分析奠定了离散化理论基础。实验表明,本文所提方法在概念格领域中可行性,离散属性数目稳定,分类精度优良。但从计算复杂度看,对于大数据应用下的高维非线性据决策连续形式背景离散,仍存在计算量偏大,实时性不佳的问题。如何在保持良好性能的同时进一步提高离散过程中决策属性的作用并提高运算速度是进一步研究的方向。

参考文献:

[1] Wile R. Restructuring lattice theory: an approach based on hierarchies of concepts [M]. Dordrecht Boston: Reidel,1982:445-470.

[2] Liu Xulong,Hong Wenxue. Using formal concept analysis to visualize relationships of syndromes intraditional Chinese medicine [J]. Medical Biometrics,LNCS,2010,6165:315-324.

[3] Cheng Chingwu, Yao Hongqing, Wu Tsungchih. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry [J]. Journal of Loss Prevention in the Process Industries,2013,26(6):1269-1278.

[4] Chen Chunhsien. Knowledge discovery using genetic algorithm for maritime situational awareness [J]. Expert Systems with Applications,2014,41(6):2742-2753.

[5] Saad F, Nürnberger A. Overview of prior-art cross-lingual information retrieval approaches [J]. World Patent Information,2012,34(4):304-314.

[6] Pera M S, Qumsiyeh R, Ng Yiukai. Web-based closed-domain data extraction on online advertisements [J]. Information Systems, 2013,38(2):183-197.

[7] 马垣. 形式概念及其新进展 [M]. 北京: 科学出版社,2010: 254-257.

(下转第 395 页)

史数据作为一个变量纳入预测模型中，形成新的多元线性回归模型，大大提高了模型用于预测流感流行的效果。

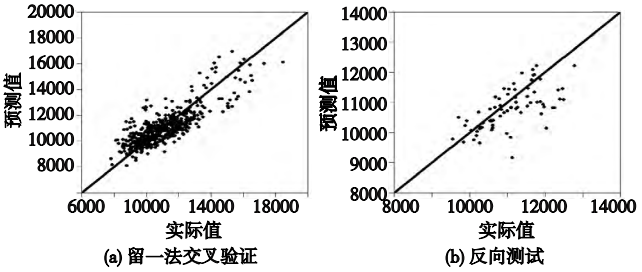


图3　整合14天前的ILI数据和百度指数搜索数据得到的模型在留一法交叉验证和反向测试中预测值和实际值的比较

中国互联网信息中心的第 33 次中国互联网发展统计报告[15]指出，截止 2013 年 12 月，中国网民数达到 6.18 亿，其中搜索引擎的用户占总网民的 79.6%，微博的网民使用率为 45.5%。百度和新浪微博作为中国互联网的佼佼者，分别在搜索引擎和微博中占有巨大市场。从网民的使用率上可以看出，百度指数统计的数据量要明显大于微指数的数据量。另一方面，新浪微博由于其受热门事件驱动的特性，会造成微指数在短期时间内的巨大波动。这两方面是最有可能造成本文中百度指数在各方面的表现优于微指数的原因。

提高时效性对于流感的防控至关重要。袁庆玉等人使用百度搜索数据能够很准确地预测流感在中国的流行，然而他们的方案是以月单位计算的，只能预测每月的整体流感流行趋势，无法细化到具体的日期，时效性不强；而本文以天为单位来统计和分析数据更具有时效性，同样也能比较准确地预测中国的流感流行。相较于黄妙玉等人利用微博来预测流感在中国的空间传播趋势，本文可以预测每日具体的 ILI 数，这对于实际的流感监测更有意义。但由于微指数只提供了 2013/3/1 之后的数据，本文统计的时间区间要和跨越的流感流行季节都要少于上述研究。所以，还需要在未来的流感监测中验证本文的模型和得到的结果。此外，在选择用于搜集百度指数和微指数数据的关键词中，本文采用了传统的先知经验选择法，相较于 Ginsberg 等人利用机器学习的方法从 5 000 万常用搜索词中筛选出 45 个最适合于流感监测的关键词，笔者的工作存在不足之处。关键词选取方案的改进将是下一步工作的重点。

4　结束语

本文作为基于互联网大数据的流感监测方案的一种探索，拥有很好的时效性。随着互联网在中国的进一步普及，基于互联网大数据的流感监测手段将发挥越来越重要的作用，它与传统的流感监测系统的结合将大大提高流感监测的效果。

参考文献：

[1] World Health Organization seasonal influenza [EB/OL]. http://www.who.int/mediacentre/factsheets/fs211/en/.

[2] Freifeld C C, Mandl K D, Reis B Y, *et al*. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports [J]. Journal of the American Medical Informatics Association, 2008, 15(2): 150-157.

[3] Polgreen P M, Chen Y, Pennock D M, *et al*. Using Internet searches for influenza surveillance [J]. Clinical Infectious Diseases, 2008, 47(11): 1443-1448.

[4] Ginsberg J, Mohebbi M H, Patel R S, *et al*. Detecting influenza epidemics using search engine query data [J]. Nature, 2008, 457(7232): 1012-1014.

[5] Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages [C]//Proc of the 1st Workshop on Social Media Analytics. [S.l.]: ACM Press, 2010: 115-122.

[6] Paul M J, Dredze M. You are what you Tweet: analyzing Twitter for public health [C]//Proc of ICWSM. 2011: 265-272.

[7] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter [C]//Proc of Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2011: 1568-1576.

[8] Achrekar H, Gandhe A, Lazarus R, *et al*. Predicting flu trends using Twitter data [C]//Proc of IEEE Conference on Computer Communications Workshops. [S.l.]: IEEE Press, 2011: 702-707.

[9] Li Jiwei, Cardie C. Early stage influenza detection from Twitter [J/OL]. Social and Information Networks, 2013, arXiv: 1309.7340.

[10] Yuan Q, Nsoesie E O, Lyu B, *et al*. Monitoring influenza epidemics in China with search query from Baidu [J]. PloS One, 2013, 8(5): e64323.

[11] Huang Jiangmiao, Zhao Hui, Zhang Jie. Detecting flu transmission by social sensor in China [C]//Proc of IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, Green Computing and Communications (GreenCom). [S.l.]: IEEE Press, 2013: 1242-1247.

[12] 百度指数平台 [DB/OL]. http://index.baidu.com.

[13] 微指数平台 [DB/OL]. http://data.weibo.com/index/hotword.

[14] 中华人民共和国国家卫生和计划生育委员会. 流感样病例暴发疫情处置指南(2012 年版) [EB/OL]. http://www.moh.gov.cn/jkj/s3577/201211/ef27b0759a014ad9951cf9dccafd9958.shtml.

[15] 中国互联网络发展状况统计报告 [EB/OL]. (2014-01). http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/P020140305346585959798.pdf.

(上接第 391 页)

[8] Chmielewski M R, Grzymala-Busse J. Global discretization of continuous attributes as preprocessing for machine learning [J]. International Journal of Approximate Reasoning, 1996, 15(4): 319-331.

[9] 黄艳，任苗苗，魏玲. 区间值决策形式背景的属性值向量约简 [J]. 计算机科学，2012，39(1)：193-197.

[10] 焦宁. 连续属性离散化算法比较研究 [D]. 合肥：合肥工业大学，2007：3-8.

[11] 王艳盼，李涛. 强协调决策形式背景的概念格属性约简 [J]. 纺织高校基础科学学报，2013，26(3)：351-354.

[12] 李金海，吕跃进. 基于概念格的决策形式背景属性约简及规则提取 [J]. 数学的实践与认识，2009，39(7)：182-188.

[13] 张涛，宋佳霖，刘旭龙，等. 基于色度学空间的多元图表示 [J]. 燕山大学学报，2010，34(2)：111-114.

[14] Janicke H, Wiebel A, Scheuermann G. Multifield visualization using local statistical of electronics [J]. IEEE Trans on Visualization & Computer Graphics, 2007, 13(6): 1384-1391.

[15] 张涛，洪文学. 基于计算几何的非线性可视化分类器设计 [J]. 电子学报，2011，39(1)：53-58.

[16] Pawlak Z, Skowron A. Rough sets and Boolean reasoning [J]. Information Sciences, 2007, 177(1): 41-73.