文章编号：2096-1472(2017)-09-34-07

基于模糊贝叶斯的改进决策方法在企业评价中的应用

冯思捷[1]，管建和[2]
(1.中航光电科技股份有限公司，河南 洛阳 471000；
2.中国地质大学(北京)，北京 100083)

摘　要：朴素贝叶斯算法是数据挖掘领域最简单的分类算法之一。为了让朴素贝叶斯能够灵活地处理连续型数据，分类过程就需要对数据进行离散化处理。而使用模糊数学理论来解决离散化问题是一个不错的选择。因此本文考虑将这两种方法结合，同时在去模糊化过程中引用了一种新型去模糊化方法(“内心法”)，从而生成一种新的模糊贝叶斯混合模型。并通过一个企业评价实例简单地验证了模糊贝叶斯算法在应对连续性数据时具有良好、可靠的分类效果。

关键词：朴素贝叶斯；模糊数学；三角模糊数；去模糊化
中图分类号：TP391　　**文献标识码**：A

The Application of the Improved Decision-Making Method Based on Fuzzy Bayes in the Enterprise Evaluation

FENG Sijie[1],GUAN Jianhe[2]
(1.*China Aviation Optical-Electrical Technology Co.,Ltd.,Luoyang* 471000,*China*;
2.*China University of Geosciences*(*Beijing*),*Beijing* 100083,*China*)

Abstract: The Naive Bayes algorithm is a simple and lucid classification way in the field of data mining.When meeting with continuous data,the algorithm usually needs to make discretization in its classifying process.Luckily,the application of relevant theories about fuzzy mathematics is a good choice to solve the discretization problem.Thus,this study decides to make a combination of the Naive Bayesian algorithm and fuzzy mathematics to generate a hybrid model and,in the meanwhile,introduces a new defuzzification method (named as *The incenter of area*) in the classification process.Through an application case of enterprise evaluation,the fuzzy Bayesian hybrid algorithm has been proved to be effective and reliable in the process of classification for continuous data.

Keywords: Naive Bayes;fuzzy math;triangular fuzzy number;defuzzification

1　引言(Introduction)

在实际生活中，某些决策型问题的处理过程通常会伴随着一定的复杂性。为了能够更好地解决这类问题，系统可以利用某些数据挖掘领域中的分类方法来得到良好且高效的决策结果。其中，朴素贝叶斯(Naive Bayes，NB)算法正是用于分类样本实例的一种简单又有效的方法。然而当它处理连续型数据时，通常的做法是利用高斯分布和极大似然估计来得到样本对应的后验概率，其计算过程往往显得较为烦琐。而由扎德提出的模糊数学理论也可以解决“连续型数据离散化”的问题。本文特将模糊数学中的三角模糊数和NB算法融合在一起，并在去模糊化过程中引入了一种新型方法。通过将构建的混合分类模型运用到企业评价应用中，体现了该模型能够具有有效且良好的分类效果。

2　朴素贝叶斯算法(Fundamentals of Naive Bayes algorithm)

朴素贝叶斯算法是最简单的一种贝叶斯分类方法，它作为一种有监督型学习方法来解决多属性分类问题。与贝叶斯信念网络相比，有研究指出朴素贝叶斯方法因其独特的“各属性间相互独立”的条件性假设而简化了整个计算过程、避免了计算带来的复杂性[1]。基于条件独立性假设和已有的先验知识，人们可以根据统计学中的贝叶斯定理学习到有用的概率信息，并最终通过计算获得的最大后验概率来获得测试样本的所属类别。

尽管独立性假设在现实生活中会显得不切实际，但是NB算法依靠它可以在很多领域根据提供的训练数据来预测出测试样本的所属类别，它通常应用于文本分类、决策预测、情

感分析等分类问题中。有研究者对朴素贝叶斯在文本分类中的应用做出了相关研究，并通过相关实验数据证实了NB算法针对小型实例数据样本集有着高精确率[2]。

定义1：(朴素贝叶斯算法)

假设给出一个样本数据集$X=\{x_1,x_2,\dots,x_k\}(i=1,2,\dots,k)$的类标号集合$C=\{C_1,C_2,\dots,C_m\}(j=1,2,\dots,m)$，还有一个描述样本属性的集合$A=\{a_1,a_2,\dots,a_n\}$，假定用来描述样本$x_i$的各属性值的每个事件$a_q(q=1,2,\dots,n)$之间相互独立。那么根据贝叶斯公式，类别$C_j$关于样本$X$的后验概率可以表示为

$$P(C_j|X)=\frac{P(x_1,x_2,\dots,x_k|C_j)\cdot P(C_j)}{P(x_1,x_2,\dots,x_k)} \quad (i=1,2,\dots,k;j=1,2,\dots,m) \tag{1}$$

其中，基于给出的条件独立假设，类别C_j下X的条件概率可以被表示为：

$P(x_1,x_2,\dots,x_k|C_j)=P(x_1|C_j)\cdot P(x_2|C_j)\cdot\dots\cdot P(x_k|C_j)=\prod_{i=1}^{k}P(x_i|C_j)$。

朴素贝叶斯分类器在决策时遵循了“最大后验法则”(the Maximum A Posterior，MAP)[3]。因此样本X的类别可以由此而得出(需要注意的是，由于$P(x_1,x_2,\dots,x_k)$是不依赖于C_j的常量，因此在下列公式中省略它)

$$NBC(X,C_j)==\underset{C_j\in C}{\arg\max}\ P(C_j)\cdot\prod_{i=1}^{k}P(x_i|C_j) \tag{2}$$

根据上面的公式可以看出，样本X的类别实际上是根据最大后验概率来得到的。需要注意的是，为了提升最终分类效果，如果在计算过程中当遇到$P(x_i|C_j)=0$的情况时，此时就需要引入“拉普拉斯标准化”(Laplace calibration)方法。也就是说，在计算$P(x_i|C_j)$的过程中对每个样本元组计数都加上1——如果对z个计数加上1的话，就必须在用于计算概率的分母上对应地加上z。关于条件概率$P(x_i|C_j)$的拉普拉斯校准公式即为

$$P(x_i|C_j)=\frac{n_t+\varepsilon}{n+z\cdot\varepsilon} \tag{3}$$

其中，n_t：在类别C_j下，事件x_i发生的样本数量；n：在所有样本实例中，类别C_j的数量；z：平滑参数，常将其设为事件x_i发生的可能取值总数(属性值x_i的种类总数)；ε：是一个值大于零的常数变量，在计算中常使其取值为1。

3　关于模糊贝叶斯的改进型决策方法(An improved decision-making model of fuzzy Naive Bayes)

人类通常在使用语言来描述描述某个事件时会伴随一些模糊现象。例如我们会用“很瘦”“比较瘦”“有点胖”或“很胖”等词语来形容一个人的体型。其中“很”“比较”和“有点”都是具有模糊性或不明确界定的词。那么模糊现象的发生也就意味着该事件存在着一定的不确定性和模糊性。

为了解决实际中遇到的模糊事件，人们尝试通过构建相关的数学模型来将不确定型变量转换成精确型变量。在1965年，Zadeh提出了一种新的数学理论——模糊数学，这种理论可以用来描述一些由人类认知或主观意识而产生的模糊事件。根据扎德提出的思想，他利用“隶属度”的概念来表示事件属于其对应模糊集合的程度，从而创建出模糊事件对应的模糊集合，并将该集合用一个特殊函数来表示[4]。其中，这个函数是由一组值域为[0,1]的隶属度组合而成的。Zadeh在他的模糊数学理论中将这个特殊函数定义为模糊事件所在域对应的隶属度函数。

定义2：(模糊集合)

假设存在一个模糊集S和论域$U=\{x_1,x_2,\dots,x_n\}$，有

$$\mu_A:U\to[0,1],\quad x\mapsto\mu_A(x)\in[0,1]$$

上述映射关系说明了在论域U中，模糊集S可以由一个函数$\mu_A(x)$来表征，而U内的每一个点都对应区间[0,1]内的某一个实数。这个函数$\mu_A(x)$通常被称作“隶属度函数”。在这个函数中，每一个函数值$\mu_A(x_i)$被看作是x_i的隶属度值。因此一个模糊集S可以按照下列公式定义：

$$S=\frac{\mu(x_1)}{x_1}+\frac{\mu(x_2)}{x_2}+\dots+\frac{\mu(x_n)}{x_n} \tag{4}$$

需要特别注意的是，$\frac{\mu(x_1)}{x_1}$的分号代表的并不是除法运算，它仅仅指出了在论域U内点x_i对应的隶属度是$\mu(x_i)$。

在多数情况下，数据或者文本样本有时因其具有主观性和不确定性而不能精确的表达内在信息。上面已经提到，隶属度是模糊数学中最基本的一个核心概念，可以通过创建一个适合的隶属度函数来表达模糊的不确定性信息。经研究者发现，通常有两种方法用来获取隶属度函数：(1)利用概率统计学和模糊数学方面的相关理论，找到一个模糊概率统计模型来表述隶属度函数；(2)可以通过模糊概率分布函数来定义一个模糊隶属度函数，比如说梯形分布、三角形分布、高斯分布等。

模糊数是模糊数学中用来表述模糊性信息的一种定量方法，它可以基于相关理论和运算方法将不确定性变量转换成精确型数值。模糊数中最常见的概念就是三角模糊数(Triangle Fuzzy Number，TFN)。三角模糊数是一种可以用来解释模糊现象、表述模糊集合的简单而高效的数学方法。它作为一种表征数据集中每个样本属性的隶属度分布的数学

模型，可以应用于多个领域用来反映出某个事件、人类语言描述或主观思想中存在的不确定性及模糊性，例如，模糊控制、模糊识别等方面。近几年有一些学者认为在一些决策系统或是评价系统中，三角模糊数可以用来表示评价权重，或是在分类问题中将其作为数学模型来用于解决决策分类问题。

实际上，三角模糊数可以看作是一个确定性和不确定性的集合体。假设一个女人测定的身高记录为160cm；这个数值可能并不是她的精确身高数值，其真实身高可能仅仅接近于、而不完全等于160cm。那么在用一个三角模糊数表征身高值时，可以用(160−x,160,160+y)来表示，其中，x和y分别是160的左、右确界。下面介绍了三角模糊数的定义。

定义3：(三角模糊数)

如果一个三角模糊数$A=\{l,m,u\}$，那么我们就可以从下列公式中获得对应的隶属度函数：

$$\mu_M(x)=\begin{cases}\dfrac{x-l}{m-l}, x\in[l,m]\\ \dfrac{x-u}{m-u}, x\in[m,u]\\ 0\ , x\in(-\infty,l]\cup[u,+\infty)\end{cases}\tag{5}$$

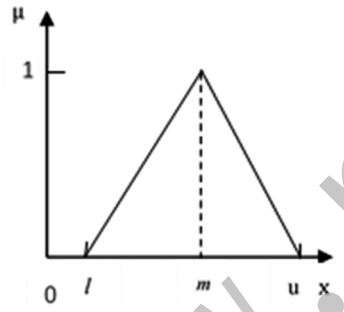


图1 三角模糊数

Fig.1 Triagular fuzzy number

根据公式(5)，我们可以称l、m和u分别是三角模糊数A的下界、中值和上界。由于中值m对应的隶属度为1，所以它对应的值是一个确定值。而除m外的位于l和u之间的值对应的隶属度值存在于区间[0,1]内。

数据挖掘领域通常会把数据分为连续型数据和离散型数据。一般情况下有以下两种形式可能会产生不确定性[5]：(1)训练数据集的类标签是由一个分布函数表示而成，这种情况可能会产生不确定性；(2)当连续型属性值以区间的形式出现时也会产生不确定性。因此在分类过程中对连续型变量进行离散化处理就显得很有必要。朴素贝叶斯算法处理的数据类型一般是离散型数据。因此当朴素贝叶斯处理的连续型数据时，就需要对其进行离散化。以往研究者们大多采用高斯分布来解决，但其计算过程一般较为烦琐，且并不能完整地解释一些由模糊现象产生的模糊性问题。因此，可以考虑将模糊数学相关理论与朴素贝叶斯算法融合在一起，可以使得生成的模糊贝叶斯混合分类器模型能够灵活地应对多种类型的数据来有效地解决分类问题。很多研究者对模糊贝叶斯问题进行了相关研究。根据Hsien−Chang Wu的研究[6]，在一些模糊环境下，贝叶斯可靠性评价系统为了简化计算过程选择将一个原始问题转换成四个子问题。之后Vibhor Kant和Kamal K.Bharadwaj[3]提出了一种基于内容的过滤方法的模糊朴素贝叶斯分类器用来解决基于相关内容的相似性问题。Kayaalp等研究学者提出了一个改进的模糊贝叶斯混合分类器用来解基于数字型数据的决策分类问题[7]。

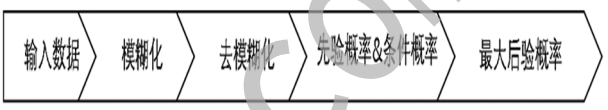


图2 模糊贝叶斯混合模型

Fig.2 The fuzzy Naive Bayes model

模糊贝叶斯算法是一种融合了模糊数学相关理论和朴素贝叶斯算法的混合模型，它在处理一些分类型问题时，可以灵活、有效地应对连续型数据。本文选择将三角模糊数和朴素贝叶斯算法进行混合，使得到的模糊贝叶斯混合分类器作为分类算法模型。这样不仅使分类过程应对不同类型的数据时的处理能力不再单一，并且还能有效地提升该过滤器的筛选能力和过滤效率。图2展示了模糊贝叶斯混合模型的搭建思路。下面介绍其操作过程：

第一步：数据准备工作。

在进入分类操作前，数据标准化过程是最主要的数据准备工作。因为不同的属性通常会存在不同的维度或具有不同的计量单位，因此有可能会影响到多属性分类问题的最终数据分析结果。那么为了消除这种潜在的不良影响，在数据准备前期对数据进行标准化处理就显得很有必要，该操作可以用来解决不同属性间的兼容性问题，从而使得他们可以存在于同一个维度解决问题。

通常情况下大多使用“最小—最大标准化方法”来对原始数据进行标准化处理。即，假设x是实数区间域内的某一个值，则经过标准化后可以得到：

$$x^*=\frac{x-min}{max-min}\tag{6}$$

其中，min和max分别是区间的最小值和最大值。

在完成数据标准化操作后，就可以开始准备创建分类模型了。假设存在一个类别集合$C=\{C_1,C_2,\dots,C_m\}$和一个样本数据集$X=\{x_1,x_2,\dots,x_n\}$，其中：每一组样本$x_i(i=1,2,\dots,n)$

都对应着某一个类别$C_p(p=1,2,\dots,m)$；而所有的x_i都可以由一个属性集合$A_i=\{A_1,A_2,\dots,A_k\}$表示。如果$x_{ij}(i=1,2,\dots,n;\ j=1,2,\dots,k)$代表了第$i$个样本的属性集$A_j$，那么根据定义4中的标准化方法，$x_{ij}$标准化后即被转换为$x'_{ij}$且该新值可以参与到接下来的分类过程中。

第二步：模糊化处理。

基于模糊集理论，这一步骤主要将属性值(经过标准化处理的)转换成它们所对应的隶属度函数。前面的内容已经提到，模糊数学的关键就是计算出数值在所处实数域内的隶属度值。因此，人们可以根据原始数据的相关特征来描述不确定型模糊信息。上面已经介绍过，通常有两种方法可以获得隶属度函数：(1)第一种方法就是利用模糊概率统计方法来解决问题；(2)第二种方法就是根据某一分布函数而专门定义一个特殊函数来描述模糊事件。有很多人尝试通过定义一个分布函数来得到隶属度函数(例如：高斯分布)或者是将不确定型变量转换成某一个特定的模糊数(例如：梯形模糊数、三角模糊数等)。那么根据人们自身定义、主观思维或者是样本数据本身的特征，就可以把语言型或者数字型数据v_{ij}转换成一个三角模糊数$(v_{ijl},v_{ijm},v_{iju})$。

第三步：去模糊化处理。

在一些理论型或者现实生活中的控制系统中，去模糊化处理是重要的一步操作，它可以将模糊数或模糊变量转换成精确的输出数据。在此之前，研究者们大多使用三角形重心或最大均值来进行去模糊化操作。但是，有研究者利用了三角形的内心提出了一种新型去模糊化方法——“内心法”(the Incentre Of Area，IOA)(注：三角形的内心就是三角形角平分线交点)[8]。那么根据定义χ(内心法定义)，就可以将三角模糊数$(x_{ijl},x_{ijm},x_{iju})$去模糊化处理后得到$x'$。关于“内心法”的定义如下：

定义5：(“内心法”去模糊化方法)

假设存在一个三角模糊数为$u=(u_1,u_2,u_3)$，而在三角模糊数的图形中的内心为$I=(I_x,I_y)$，即有

$$I_x=\frac{u_1\alpha+u_2\beta+u_3\gamma}{\alpha+\beta+\gamma} \tag{7}$$

$$I_y=\frac{\beta}{\alpha+\beta+\gamma} \tag{8}$$

其中，$\alpha=\sqrt{(u_3-u_2)^2+1}$，$\beta=u_3-u_1$，$\gamma=\sqrt{(u_2-u_1)^2+1}$

因此，三角模糊数u对应的精确值就是I_x，而对应的隶属度值是I_y。

第四步：先验概率和条件概率。

根据朴素贝叶斯算法的基本概念可知，先验概率和条件概率可以由下列公式求得

$$P(C_p)=\frac{\text{类别为}C_p\text{的样本数目}}{\text{样本总数}} \tag{9}$$

$$P(x_i|C_p)=\frac{\text{类别为}C_p\text{时，属性}A_j\text{下值为}x_i\text{的样本数目}+1}{\text{类别为}C_p\text{的样本总数}+|V|} \tag{10}$$

其中，公式(10)使用了拉普拉斯校准，而变量V代表了类别A_j的取值种类数。

第五步：最大后验概率。

依据提供的训练数据样本的相关数据值，由公式(9)、公式(10)求得的先验概率和条件概率。然后参考最大后验概率法则，见式(2)，就可以对测试样本数据计算、分析出测试用例的最终分类结果。

4 关于企业评价的简单实例应用(An example of the application on commercial enterprise evaluation)

通常专家会设定出专门的评价标准来对不同的企业进行评估，以此将企业划分为不同的类型。然而，不管所用的评价打分是数值型还是文本型，专家给出的评价值有时仍可能会存在着模糊性。在语言评价系统中可以通过将语言评价值转换成模糊数这个方法来进行分类[9]。根据这种思路，本文将构建的模糊贝叶斯混合模型应用到企业评价中，具体过程如下。

4.1 数据准备

在对企业评估的过程中，专家会根据相关专业知识或者自己的经验而专门设定评分规则来对企业进行打分，并最终将企业划分成三种类别(分别为Ⅰ、Ⅱ和Ⅲ)。在给出的企业评价样本集中，一共考察了四种属性，如表1所示。

表1 四个属性的值域区间

Tab.1 The range value of four attributes

属性	值域最小值	值域最大值
财产效益	4.3	7.9
资产营运	2.0	4.4
偿债能力	1.0	6.9
发展能力	0.1	2.5

根据表1提供的数据，可以将每一个属性的值域区间依次划分成三个子区间。为了使专家能够对每个企业的属性指标做出评价，特设定两个人工语言评价集：(1)有关“财产效益”和“偿债能力”的语言评价集合：A={L,M,H}(其中，L、M和H分别代表低、中等和高)；(2)有关“资产营运”和“发展能力”的语言评价集：B={W,M,S}(其中，W、M和S分别代表弱、中等和强)。这样，评价集A、B中的每一个元素

(即人工语言评价值)就可以分别被用来定义经过划分得到的属性值域子区间。图3展示了四个属性的值域划分结果，以及每个子区间对应的评价值。

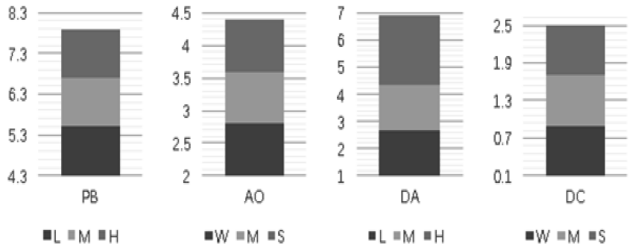


图3 每个子区间及其对应的评价值

Fig.3 Every subinterval and its corresponding evaluation value

企业评价原始数据集描述了每个训练样本的相关数据及对应的专家评价值，详见表2，即每一个属性由两个子属性表示：获得的专家打分(表2中的“d”列)和相对应的语言评价(表2中的“v”)列。

表2 原始数据信息和相对应的人工语言评价值

Tab.2 Raw data information and the corresponding artificial language evaluation value

实例	财产效益		资产营运		偿债能力		发展能力		类别
	d	v	d	v	d	v	d	v	
1	5.8	M	2.7	W	5.1	H	1.9	S	Ⅲ
2	5	L	2	W	3.5	M	1	M	Ⅱ
3	5	L	2.3	W	3.3	M	1	M	Ⅱ
4	6.3	M	3.3	M	6	H	2.5	S	Ⅲ
5	4.9	L	2.4	W	3.3	M	1	M	Ⅱ
6	7.1	H	3	M	5.9	H	2.1	S	Ⅲ
7	6.7	H	2.5	W	5.8	H	1.8	S	Ⅲ
8	5.1	L	3.8	S	1.9	L	0.4	W	Ⅰ
9	4.7	L	3.2	M	1.3	L	0.2	W	Ⅰ
10	5.9	M	3.2	M	4.8	H	1.8	S	Ⅱ
11	6.9	H	3.1	M	4.9	H	1.5	M	Ⅱ
12	7.7	H	2.6	W	6.9	H	2.3	S	Ⅲ
13	6.3	M	3.4	M	5.6	H	2.4	S	Ⅲ
14	4.9	L	3	M	1.4	L	0.2	W	Ⅰ
15	5	L	3.5	M	1.6	L	0.6	W	Ⅰ
16	5.8	M	4	S	1.2	L	0.2	W	Ⅰ
17	5.1	L	3.5	M	1.4	L	0.2	W	Ⅰ
18	6.4	M	3.2	M	4.5	H	1.5	M	Ⅱ
19	7	H	3.2	M	4.7	H	1.4	M	Ⅱ
20	4.5	L	2.3	W	1.3	L	0.3	W	Ⅰ
21	7.7	H	3.8	S	6.7	H	2.2	S	Ⅲ

4.2 数据标准化

准备好分类所需的训练样本数据后，接下来还需要对这些数据进行预处理操作。虽然根据一些已设定好的打分规则，就可以获得专家对企业样本的打分及其对应的语言评价值。但考虑到不同专家存在不同的主观思想来进行打分，且不同的属性存在有不同的取值区间(表1)，因此需要根据公式(6)来将表2中的原始数据进行标准化处理。经过标准化处理后所得的数据详见表3。

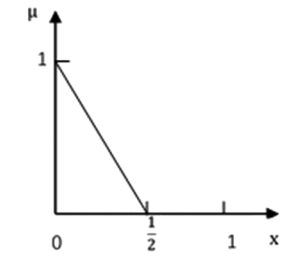
表3 标准化数据表

Tab.3 The standardized dataset

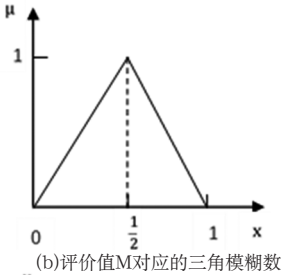
实例	财产效益		资产营运		偿债能力		发展能力		类别
	d	v	d	v	d	v	d	v	
1	0.4167	M	0.2917	W	0.6949	H	0.75	S	Ⅲ
2	0.1944	L	0	W	0.4237	M	0.375	M	Ⅱ
3	0.1944	L	0.125	W	0.3898	M	0.375	M	Ⅱ
4	0.5556	M	0.5417	M	0.8475	H	1	S	Ⅲ
5	0.1667	L	0.1667	W	0.3898	M	0.375	M	Ⅱ
6	0.7778	H	0.4167	M	0.8305	H	0.8333	S	Ⅲ
7	0.6667	H	0.2083	W	0.8136	H	0.7083	S	Ⅲ
8	0.2222	L	0.75	S	0.1525	L	0.125	W	Ⅰ
9	0.1111	L	0.5	M	0.0508	L	0.0417	W	Ⅰ
10	0.4444	M	0.5	M	0.6441	H	0.7083	S	Ⅱ
11	0.7222	H	0.4583	M	0.661	H	0.5833	M	Ⅱ
12	0.9444	H	0.25	W	1	H	0.9167	S	Ⅲ
13	0.5556	M	0.5833	M	0.7797	H	0.9583	S	Ⅲ
14	0.1667	L	0.4167	M	0.0678	L	0.0417	W	Ⅰ
15	0.1944	L	0.625	M	0.1017	L	0.2083	W	Ⅰ
16	0.4167	M	0.8333	S	0.0339	L	0.0417	W	Ⅰ
17	0.2222	L	0.625	M	0.0678	L	0.0417	W	Ⅰ
18	0.5833	M	0.5	M	0.6271	H	0.5417	M	Ⅱ
19	0.75	H	0.5	M	0.6271	H	0.5417	M	Ⅱ
20	0.0556	L	0.125	W	0.0508	L	0.0833	W	Ⅰ
21	0.9444	H	0.75	S	0.9661	H	0.875	S	Ⅲ

4.3 使用模糊化得到的评价值分隔经过标准化处理的数据

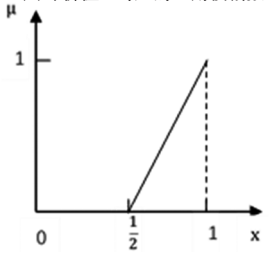
在完成数据标准化操作后，可以考虑将专家打分对应的语言评价值(即语言评价集合A和集合B中的每个元素)转换成不同的三角模糊数。假定存在一个语言变量集合$I=\{i_1,i_2,\dots,i_m,\dots,i_n\}$，该集合由一组有序的语言值组合而成，其中$i_m(m=1,2,\dots,n)$是集合$I$中的某一个语言评价值。那么可以将变量i_m定义成一个三角模糊数$(\frac{m-1}{n},\frac{m}{n},\frac{m+1}{n})$。有关评价集合元素L、M、H、W和S的三角模糊数如图4所示。



(a)评价值L和W对应的三角模糊数



(b)评价值M对应的三角模糊数



(c)评价值W和S对应的三角模糊数

图4 每个评价值对应的三角模糊数

Fig.4 The triangular fuzzy number of every evaluation value

接下来参考“内心法”，对人工语言评价值(L、M、H、W和S)对应的三角模糊数(f_1,f_2,f_3)进行去模糊化处理，这样就可以求出其对应的精确值(也就是下面计算过程中的I_{x1}、I_{x2}和I_{x3})。具体的计算过程如下：

(1)L和W对应的三角模糊数为$f_1=(0,0,\frac{1}{2})$：

即有：$\alpha_1=\sqrt{\left(\frac{1}{2}-0\right)^2+1}=\frac{\sqrt{5}}{2},\beta_1=\frac{1}{2},\ \gamma_1=\sqrt{(0-0)^2+1}=1$

$\therefore I_{x1}=0.1910$

(2)M对应的三角模糊数为$f_2=(0,\frac{1}{2},1)$：

即有：$\alpha_2=\sqrt{\left(\frac{1}{2}-0\right)^2+1}=\frac{\sqrt{5}}{2},\ \beta_1=1,\gamma_1=\sqrt{\left(\frac{1}{2}-0\right)^2+1}=\frac{\sqrt{5}}{2}$

$\therefore I_{x2}=0.5$

(3)W和S对应的三角模糊数为$f_3=(\frac{1}{2},1,1)$：

即有：$\alpha_1=\sqrt{(1-1)^2+1}=1,\ \beta_1=\frac{1}{2},\ \gamma_1=\sqrt{\left(\frac{1}{2}-0\right)^2+1}=\frac{\sqrt{5}}{2}$

$\therefore I_{x3}=0.8090$

由于标准化处理后得到的数据都位于区间[0,1]内，可以将I_{x1}、I_{x2}和I_{x3}作为划分[0,1]区间的分隔值，就能得到如表4所示的四个子区间。然后让得到的四个子区间分别对应一个新属性值(即A、B、C或D)，这步操作即完成连续数据离散化处理。

表4 四个离散型新属性值

Tab.4 Four discrete new attribute values

被划分的子区间	对应的新属性值
$[0,I_{x1}]$	A
$(I_{x1},I_{x2}]$	B
$(I_{x2},I_{x3}]$	C
$(I_{x3},1]$	D

4.4 用例测试

将表3里每个标准化后得到的取值按照表4中的对应区间找到对应的新属性值，如表5所示，就可以实现“将连续型数据离散化”的目的。

表5 实现离散化处理后得到的新数据

Fig.5 The new acquired data after discretization

实例	财产效益	资产营运	偿债能力	发展能力	类别
1	B	B	C	C	Ⅲ
2	B	A	B	B	Ⅱ
3	B	A	B	B	Ⅱ
4	C	C	D	D	Ⅲ
5	A	A	B	B	Ⅱ
6	C	B	D	D	Ⅲ
7	C	B	D	C	Ⅲ
8	B	C	A	A	Ⅰ
9	A	B	A	A	Ⅰ
10	B	B	C	C	Ⅱ
11	C	B	C	C	Ⅱ
12	D	B	D	D	Ⅲ
13	C	C	C	D	Ⅲ
14	A	B	A	A	Ⅰ
15	B	C	A	B	Ⅰ
16	B	D	A	A	Ⅰ
17	B	C	A	A	Ⅰ
18	C	B	C	C	Ⅱ
19	C	B	C	C	Ⅱ
20	A	A	A	A	Ⅰ
21	D	C	D	D	Ⅲ

完成离散化处理后就可以将得到的新属性值运用于朴素贝叶斯算法中，并计算出测试实例的所属类别。为了验证并测试上面提出的混合算法模型，特设定如下六个测试用例

$\{T_1, T_2, \cdots, T_6\}$。(注："$pb$" "$ao$" "$da$" "$dc$"是四个属性指标("财产效益" "资产营运" "偿债能力" "发展能力")的英文缩写)。

$$T_1 = \{pb = 5.7, ao = 3.8, da = 1.7, dc = 0.3\}$$
$$T_2 = \{pb = 6.8, ao = 2.8, da = 4.8, dc = 1.4\}$$
$$T_3 = \{pb = 5.5, ao = 2.5, da = 4.0, dc = 1.3\}$$
$$T_4 = \{pb = 5.0, ao = 3.4, da = 1.6, dc = 0.4\}$$
$$T_5 = \{pb = 6.7, ao = 3.0, da = 5.2, dc = 2.3\}$$
$$T_6 = \{pb = 6.3, ao = 3.3, da = 4.7, dc = 1.6\}$$

下面专门列出其中两个测试用例T_1和T_2的具体计算过程。

(1)T_1用例:

根据公式(9),可以求出三种类别的先验概率:

$$\mathrm{P}(\mathrm{I}) = \frac{7}{21} = \frac{1}{3},\ \mathrm{P}(\mathrm{II}) = \frac{7}{21} = \frac{1}{3},\ \mathrm{P}(\mathrm{III}) = \frac{7}{21} = \frac{1}{3}$$

接下来根据公式(6)对T_1的属性值进行标准化处理:

pb'=0.3889, ao'=0.75, da'=0.1186, dc'=0.0833

根据表2—表4可知,T_1就可以被定义为$\{pb = B, ao = C, da = A, dc = A\}$。

要想获得T_1的所属类别,需要计算出条件概率和最大后验概率。

为了保证分类时的计算精准率,在计算条件概率的过程中需要对其进行拉普拉斯校准(公式(10))。

i. $\mathrm{P}(pb = \mathrm{B}|\mathrm{I}) = \frac{4+1}{7+4} = \frac{5}{11},\ \mathrm{P}(ao = \mathrm{C}|\mathrm{I}) = \frac{3+1}{7+4} = \frac{4}{11}$

$\mathrm{P}(da = \mathrm{B}|\mathrm{I}) = \frac{7+1}{7+4} = \frac{8}{11},\ \mathrm{P}(dc = \mathrm{A}|\mathrm{I}) = \frac{6+1}{7+4} = \frac{7}{11}$

$$\therefore NBC(T_1, \mathrm{I}) = \frac{5}{11} \times \frac{4}{11} \times \frac{8}{11} \times \frac{7}{11} \times \frac{1}{3} = 0.0225$$

ii. $\mathrm{P}(pb = \mathrm{B}|\mathrm{II}) = \frac{3+1}{7+4} = \frac{4}{11},\ \mathrm{P}(ao = \mathrm{C}|\mathrm{II}) = \frac{0+1}{7+4} = \frac{1}{11}$

$\mathrm{P}(da = \mathrm{A}|\mathrm{II}) = \frac{0+1}{7+4} = \frac{1}{11},\ \mathrm{P}(dc = \mathrm{A}|\mathrm{II}) = \frac{0+1}{7+4} = \frac{1}{11}$

$$\therefore NBC(T_1, \mathrm{II}) = \frac{4}{11} \times \frac{1}{11} \times \frac{1}{11} \times \frac{1}{11} \times \frac{1}{3} = 0.0001$$

iii. $\mathrm{P}(pb = \mathrm{B}|\mathrm{III}) = \frac{1+1}{7+4} = \frac{2}{11},\ \mathrm{P}(ao = \mathrm{C}|\mathrm{III}) = \frac{3+1}{7+4} = \frac{4}{11}$

$\mathrm{P}(da = \mathrm{A}|\mathrm{III}) = \frac{0+1}{7+4} = \frac{1}{11},\ \mathrm{P}(dc = \mathrm{A}|\mathrm{III}) = \frac{0+1}{7+4} = \frac{1}{11}$

$$\therefore NBC(T_1, \mathrm{III}) = \frac{2}{11} \times \frac{4}{11} \times \frac{1}{11} \times \frac{1}{11} \times \frac{1}{3} = 0.0002$$

那么根据MAP(最大后验法则,公式(2)),综上可得:T_1属于类别Ⅰ。

(2)T_2用例:

在经过标准化处理后,T_2可以被定义为$\{pb = C, ao = B, da = C, dc = C\}$。

i. $\mathrm{P}(pb = \mathrm{C}|\mathrm{I}) = \frac{0+1}{7+4} = \frac{1}{11},\ \mathrm{P}(ao = \mathrm{B}|\mathrm{I}) = \frac{2+1}{7+4} = \frac{3}{11}$

$\mathrm{P}(da = \mathrm{C}|\mathrm{I}) = \frac{0+1}{7+4} = \frac{1}{11},\ \mathrm{P}(dc = \mathrm{C}|\mathrm{I}) = \frac{0+1}{7+4} = \frac{1}{11}$

$$\therefore NBC(T_2, \mathrm{I}) = \frac{1}{11} \times \frac{3}{11} \times \frac{1}{11} \times \frac{1}{11} \times \frac{1}{3} = 0.0001$$

ii. $\mathrm{P}(pb = \mathrm{C}|\mathrm{II}) = \frac{3+1}{7+4} = \frac{4}{11},\ \mathrm{P}(ao = \mathrm{B}|\mathrm{II}) = \frac{4+1}{7+4} = \frac{5}{11}$

$\mathrm{P}(da = \mathrm{C}|\mathrm{II}) = \frac{4+1}{7+4} = \frac{5}{11},\ \mathrm{P}(dc = \mathrm{C}|\mathrm{II}) = \frac{4+1}{7+4} = \frac{5}{11}$

$$\therefore NBC(T_2, \mathrm{II}) = \frac{4}{11} \times \frac{5}{11} \times \frac{5}{11} \times \frac{5}{11} \times \frac{1}{3} = 0.0114$$

iii. $\mathrm{P}(pb = \mathrm{C}|\mathrm{III}) = \frac{4+1}{7+4} = \frac{5}{11},\ \mathrm{P}(ao = \mathrm{B}|\mathrm{III}) = \frac{4+1}{7+4} = \frac{5}{11}$

$\mathrm{P}(da = \mathrm{C}|\mathrm{III}) = \frac{2+1}{7+4} = \frac{3}{11},\ \mathrm{P}(dc = \mathrm{C}|\mathrm{III}) = \frac{2+1}{7+4} = \frac{3}{11}$

$$\therefore NBC(T_2, \mathrm{III}) = \frac{5}{11} \times \frac{5}{11} \times \frac{3}{11} \times \frac{3}{11} \times \frac{1}{3} = 0.0051$$

综上所述可知,可以看出样例T_2属于类别Ⅱ。

在上述应用模糊贝叶斯混合模型的简单实例中,通过提供一些企业样本用例可以测试出该混合算法模型的分类性能。可以看出:结合了"内心法"创建的混合模型实现了将连续型数据实例离散化的目标,使朴素贝叶斯分类算法在处理连续型数据时的计算过程变得更为灵活,从而使得该模型能够有效地获得实例的所属类别。

5 结论(Conclusion)

在数据挖掘领域,研究者们常常会将模糊数学和分类算法进行结合,在分类过程中按照"模糊化—去模糊化"的模式来对数据进行处理。在以往的研究中,人们大多采用COA方法和MOM方法进行去模糊化操作。为了改善模糊贝叶斯混合算法,本文尝试将一种新型去模糊化方法("内心法")融入朴素贝叶斯算法中得到一个混合分类模型。在企业评价简单实例应用中,可以看到模糊贝叶斯混合分类模型不仅实现了对连续型数据离散化的目标,而且使得数据能够更好地参与朴素贝叶斯算法的分类过程中。然而本次试验中用于测试的实验用例数量并不十分充足,因此在今后的研究学习中需要继续增加测试样本数量,以进一步提升该模糊贝叶斯混合分类器的分类性能。

参考文献(References)

[1] Jiang L,et al.Structure extended multinomial Naive Bayes[J]. Information Sciences,2016,329(C):346−356.

[2] Lei L I,Huang Y G,Liu Z W.Chinese text classification for small sample set[J].Journal of China Universities of Posts & Telecom munications,2011,18:83−89.

[3] Kant V,Bharadwaj K K.Integrating Collaborative and Reclusive Methods for Effective Recommendations:A Fuzzy Bayesian Approach[J].International Journal of Intelligent

Systems,2013,28(11):1099−1123.

[4] Zadeh L A.Fuzzy sets[C].Fuzzy Sets,Fuzzy Logic & Fuzzy Systems.World Scientific Publishing Co.Inc.1996:394−432.

[5] Bounhas M,et al.Naive possibilistic classifiers for imprecise or uncertain numerical data[J].Fuzzy Sets & Systems,2014,239(1):137−156.

[6] Wu H C.Bayesian system reliability assessment under fuzzy environments[J]. Reliability Engineering & System Safety,2004,83(3):277−286.

[7] Kayaalp N.An Aggregated Fuzzy Naive Bayes Data Classifier[M]. Elsevier Science Publishers B.V.2015.

[8] Rouhparvar H,Panahi A.A new definition for defuzzification of generalized fuzzy numbers and its application[M].Elsevier Science Publishers B.V.2015.

[9] Wang J,et al.A synthetic method for knowledge management performance evaluation based on triangular fuzzy number and group support systems[J].Applied Soft Computing,2016, 39(C):11−20.

作者简介：

冯思捷(1992−)，女，硕士，技术员.研究领域：数据挖掘.

管建和(1962−)，男，博士，教授.研究领域：数据挖掘.

（上接第44页）

的用户服务，增强产品的市场竞争力。

参考文献(References)

[1] 洪利,等.无线CPU与移动IP网络开发技术[M].北京:北京航空航天大学出版社,2015.

[2] W.Richard Stevers(美).TCP/IP详解(第三卷)协议[M].北京:北京大学出版社,2015.

[3] WM_Q2686_modules_spec_sheet [S].Wavecom Corporation, 2016.

[4] ADL_User_Guide For OpenAt@OS[S].Wavecom Corporation, 2016.

[5] 胡静静.实现基于GPRS的无线远程IAP功能[J].单片机与嵌入式系统应用,2005(1):21−23.

作者简介：

鲍海森(1977−)，男，本科，高级工程师.研究领域：车联网技术，战略方向规划，车联网商业模式分析.

（上接第48页）

法在运行600轮左右网络死亡，显著延长的系统的生命周期。

5　结论(Conclusion)

本文针对无线水质监测系统中监测节点的能量空洞现象造成网络过早死亡的问题。从网络路径优化的角度，采用量子遗传算法进行数据传输路径的优化，解决水质监测节点的过早死亡问题，主要表现在：

(1)优化了无线水质监测网络的数据传输路径，采用改进量子遗传算法，提高路径搜寻的成功率，避免收敛到局部最优解；且能够在较少的代数内找到最优路径，路径优化成功率高。

(2)充分考虑各节点的传输能耗和剩余能量等因素，选择最合适的通信路径，避免能量空洞现象过早出现，有效延长了网络生存周期。

参考文献(References)

[1] 雷霖,李伟峰,王厚军.基于遗传算法的无线传感器网络路径优化[J].电子科技大学学报,2009,38(2):227−230.

[2] 童孟军,关华丞.基于蚁群算法的能量均衡多路径路由算法的研究[J].传感技术学报,2013,26(3):425−434.

[3] Zhu X,Zhang Y.Wireless sensor network path optimization based on particle swarm algorithm[J].Computer Engineering,2010,3(4):534−537.

[4] 唐义龙,等.基于量子遗传算法的无线传感器网络路由研究[J].传感器与微系统,2011,30(12):68−70.

[5] 钱晓华,王俊平.基于量子遗传算法的无线传感器网络路由[J].辽宁大学学报(自然科学版),2010,37(2):113−115.

[6] 邹少军.基于量子遗传算法的无线传感器网络路径优化[J].计算机测量与控制,2010,18(3):723−726.

[7] 夏俊,等.基于量子遗传算法的无线传感网络路由优化[J].同济大学学报(自然科学版),2015,43(7):1097−1103.

[8] Sharma R.Energy Holes Avoiding Techniques in Sensor Networks:A survey[J].2015,20(4):204−208.

[9] Estrin D,Srivastava M,Sayeed A.Tutorial on Wireless Sensor Networks[J].Technologies Protocols & Applications, 2002,13(4):317−328(12).

[10] Xu Yulong,Wang Xiaopeng,Zhang Han.Comparative study on the optimal path problem of wireless sensor networks[C].2016 IEEE International Conference on Mechatronics and Automation,2016:2234−2239.

作者简介：

申庆祥(1990−)，男，硕士生.研究领域：无线传感器网络，物联网技术.

张宇华(1975−)，男，博士,副教授.研究领域：无线传感器网络，能源管理与节能治理.