混合型数据的邻域条件互信息熵属性约简算法

兰海波
中国气象局公共气象服务中心，北京 100081

摘要
属性约简是粗糙集理论的重要研究内容之一，其主要目的是消除信息系统中不相关的属性，降低数据维度并提高数据知识发现性能。然而，基于粗糙集的属性约简方法大多没有考虑属性之间的依赖性，使得最终的属性约简结果存在一定的冗余属性。对此，提出一种基于邻域条件互信息熵的属性约简算法。首先，在传统邻域熵的基础上，针对混合型数据，提出混合型邻域互信息熵模型和混合型邻域条件互信息熵模型；然后利用这两种熵模型进行混合型信息系统的属性依赖度评估和属性启发式搜索，并设计出一种属性约简算法；最后通过UCI数据集的实验分析，证明了提出的算法具有较高的属性约简性能。

关键词
粗糙集；属性约简；邻域；互信息熵；条件互信息熵

中图分类号：TP18　　文献标志码：A　　doi: 10.11959/j.issn.2096-0271.2022066

Neighborhood conditional mutual information entropy attribute reduction algorithm for hybrid data

LAN Haibo
CMA Public Meteorological Service Centre, Beijing 100081, China

Abstract
Attribute reduction is an important research content of the rough set theory. Its main purpose is to eliminate irrelevant attributes in information systems, reduce data dimensions and improve data knowledge discovery performance. However, most of the attribute reduction methods based on a rough set do not consider the dependence between attributes, which makes the final attribute reduction result have some redundant attributes. An attribute reduction algorithm based on neighborhood conditional mutual information entropy was proposed. Firstly, based on the traditional neighborhood entropy, a hybrid neighborhood mutual information entropy model and a hybrid neighborhood conditional mutual information entropy model were proposed for hybrid data. Then, the two entropy models were used to evaluate the attribute dependence and attribute heuristic search of the hybrid information system,

and an attribute reduction algorithm was designed. Finally, through the experimental analysis of UCI data sets, it was proved that the algorithm had higher attribute reduction performance.

Key words
rough set, attribute reduction, neighborhood, mutual information entropy, conditional mutual information entropy

0 引言

在大数据应用情景下，具有噪声、无关或冗余特征的数据集对数据挖掘、知识发现和模式识别产生了巨大的挑战[1-2]。如何从数据集所有属性中选择出最优属性子集是各种学习任务的重要研究课题。属性约简是粗糙集理论的重要研究分支，其主要目的是消除信息系统中不相关的属性，降低数据维度并提高数据知识发现性能[3]。

基于粗糙集理论，学者们提出了多种属性约简算法。例如，Hu Q H等人[4]基于邻域粗糙集，将邻域依赖度作为数值型信息系统的属性评估，提出一种属性约简算法；Pang Q Q等人[5]提出一种基于邻域区分度的半监督属性约简算法；在Pang Q Q等人的基础上，Hu M等人[6]在邻域粗糙集下提出权重邻域依赖度，并构造一种改进的属性约简算法；Shu W H等人[7]对邻域粗糙集进行增量式构造，提出一种高效的增量式属性约简算法；盛魁等人[8]对邻域区分度进行增量式构造，提出一种新的属性约简算法；姚晟等人[9]将这些属性约简算法进一步拓展，提出非平衡数据下不完备混合型信息系统的属性约简算法。另外，部分学者采用其他类型的粗糙集模型进行属性约简算法的设计，例如，Wang C Z等人[10]在模糊粗糙集下提出自信息，并进行属性约简算法的设计；Yuan Z等人[11]利用模糊粗糙集提出混合型数据的非监督属性约简算法；栾雨雨等人[12]利用混沌离散粒子群提出一种新的粗糙集属性约简算法；Hu M等人[13]利用K近邻粗糙集模型提出一种新颖的属性约简算法；桑彬彬等人[14]利用优势粗糙集构造出一种属性约简算法。

利用互信息熵进行属性约简近年来受到了学者们越来越多的关注。熊菊霞等人[15]提出邻域互信息熵的混合型数据属性约简算法，陈帅等人[16-17]提出邻域互补信息度量的属性约简算法，姚晟等人[18]提出邻域互信息熵的非单调性属性约简算法。然而，这些属性约简算法大多没有考虑属性之间的相互作用，即在进行属性约简的搜索过程中，选择重要度高的属性作为候选属性，而没有考虑所选属性的独立性，新选择的属性与已有的属性可能存在一定的依赖关系，这使得最终的属性约简结果可能存在一定的冗余性[19]。互信息熵与条件互信息熵是评估随机变量独立性的一种重要度量方法[15]，本文将利用这两种度量方法提出一种新的属性约简算法。同时，实际应用环境下的数据集往往是数值型和离散型混合类型，例如对于医疗信息系统，患者的性别、听觉、视觉、嗅觉等都是离散型的属性，身高、体重和血液检查中各种酶的指标都是数值型的属性，因此本文将研究混合型信息系统下的属性约简问题。

首先，本文在邻域粗糙集模型的基础上，构造出混合型信息系统下的邻域信息熵模型，并进一步提出混合型邻域互信息熵模型和混合型邻域条件互信息熵模型；然后，将提出的混合型邻域互信息熵和混合型邻域条件互信息熵用于混合型信息系统属性之间的相关性度量；最后，将这两种熵度量作为启发式函数设计出一种属性约简算法，并通过6个UCI数据集的属性约简实验，

证明了本文的属性约简算法通过考虑属性之间的依赖性可以提高约简结果的分类性能，同时本文算法也具有较小的属性约简耗时。

1 基本理论

将邻域信息系统表示为二元组 $\mathrm{NIS}=(U,\mathrm{AT}=C\cup D)$，其中，$U=\{x_1,x_2,\cdots,x_n\}$ 是一个非空有限对象或样本的集合，称之为论域；$\mathrm{AT}=C\cup D$ 是一个非空有限属性或特征的集合，称之为属性全集，其包含两个部分，分别称之为条件属性集 C 和决策属性集 D。

在邻域信息系统 $\mathrm{NIS}=(U,\mathrm{AT}=C\cup D)$ 中，通常使用距离度量来评估信息系统中对象之间的相似性，对于属性子集 $A=\{a_1,a_1,\cdots,a_m\}\subseteq C$，对象 $x,y\in U$ 的距离度量一般被定义为：

$$\Delta_A(x,y)=\left(\sum_{i=1}^{m}|a_i(x)-a_i(y)|^{\lambda}\right)^{\frac{1}{\lambda}} \quad (1)$$

其中，$a_i(x)$ 表示对象 x 在属性 a_i 下的属性值，$a_i(y)$ 表示对象 y 在属性 a_i 下的属性值，λ 的取值范围一般为 $(1,2,\cdots,+\infty)$。基于该度量函数，可以在邻域信息系统下构造出邻域关系。

定义1[4]：设邻域信息系统表示为 $\mathrm{NIS}=(U,\mathrm{AT}=C\cup D)$，则属性子集 $A\subseteq C$ 确定的邻域关系如下。

$$\mathrm{NR}^{\delta}(A)=\{(x,y)\in U\times U\mid \Delta_A(x,y)\leqslant\delta\} \quad (2)$$

其中，δ 被称为邻域关系的邻域半径。邻域关系满足自反性和对称性，但不一定满足传递性。利用邻域关系可以得到邻域信息系统中每个对象的邻域类 $\delta_A(x)$：

$$\delta_A(x)=\{y\in U\mid (x,y)\in \mathrm{NR}^{\delta}(A)\} \quad (3)$$

定义2[4]：设邻域信息系统表示为 $\mathrm{NIS}=(U,\mathrm{AT}=C\cup D)$，属性子集 $A\subseteq C$ 确定的邻域关系为 $\mathrm{NR}^{\delta}(A)$，则对象集 $X\subseteq U$ 在邻域关系 $\mathrm{NR}^{\delta}(A)$ 下的邻域下近似集 $\underline{\mathrm{Apr}}_A^{\delta}(X)$ 和邻域上近似集 $\overline{\mathrm{Apr}}_A^{\delta}(X)$ 分别定义如下。

$$\underline{\mathrm{Apr}}_A^{\delta}(X)=\{x\in U\mid \delta_A(x)\subseteq X\} \quad (4)$$

$$\overline{\mathrm{Apr}}_A^{\delta}(X)=\{x\in U\mid \delta_A(x)\cap X\neq\varnothing\} \quad (5)$$

信息熵模型是评估信息系统不确定性的一种重要方法，Hu Q H等人[20]在邻域信息系统下提出了一种邻域熵模型。

定义3[20]：设邻域信息系统表示为 $\mathrm{NIS}=(U,\mathrm{AT}=C\cup D)$，属性子集 $A\subseteq C$ 确定的邻域关系为 $\mathrm{NR}^{\delta}(A)$，对象 $x\in U$ 在 $\mathrm{NR}^{\delta}(A)$ 下的邻域类为 $\delta_A(x)$，那么邻域关系 $\mathrm{NR}^{\delta}(A)$ 确定的邻域熵 $\mathrm{NE}^{\delta}(A)$ 定义如下。

$$\mathrm{NE}^{\delta}(A)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A(x_i)|}{|U|} \quad (6)$$

Hu Q H等人[20]提出的邻域熵模型在邻域粗糙集的不确定性度量和属性约简方面发挥了重要作用，使得邻域熵模型成为邻域粗糙集的重要研究内容。

2 混合型信息系统的邻域条件互信息熵模型

然而，实际应用中的数据包含数值型和标记型，传统的邻域粗糙集模型仅适用于数值型，针对这一局限性，盛魁等人[8]提出了基于混合型信息系统的邻域粗糙集模型。

定义4[8]：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，其中 $C=C_n\cup C_m$ 且 $C_n\cap C_m=\varnothing$，C_n 为条件属性集中的数值型属性子集，C_m 为条件属性集中的标记型属性子集。对于 $A=A_n\cup A_m$，其中 $A_n\subseteq C_n$、$A_m\subseteq C_m$，那么 $A\subseteq C$ 确定的混合邻域关系如下。

$$\begin{aligned}&\mathrm{MNR}^{\delta}(A)=\{(x,y)\in\\&U\times U\mid \Delta_{A_n}(x,y)\leqslant\\&\delta\wedge(\forall a\in A_m,a(x)=a(y))\}\end{aligned} \quad (7)$$

同时，对于 $\forall x \in U$ ，在混合邻域关系 $\mathrm{MNR}^{\delta}(A)$ 下的邻域类 $\delta_A^*(x)$ 定义为：

$$\delta_A^*(x)=\{y\in U\mid (x,y)\in \mathrm{MNR}^{\delta}(A)\} \quad (8)$$

基于混合信息系统的混合邻域关系和邻域类，盛魁等人[8]进一步提出了一种改进的邻域粗糙集模型。

定义5[8]：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ， $A=A_n\cup A_m$ 确定的混合邻域关系为 $\mathrm{MNR}^{\delta}(A)$ ，那么对象集 $X\subseteq U$ 在 $\mathrm{MNR}^{\delta}(A)$ 下的邻域下近似集 $\underline{\mathrm{MApr}}_A^{\delta}(X)$ 和邻域上近似集 $\overline{\mathrm{MApr}}_A^{\delta}(X)$ 分别定义如下。

$$\underline{\mathrm{MApr}}_A^{\delta}(X)=\{x\in U\mid \delta_A^*(x)\subseteq X\} \quad (9)$$

$$\overline{\mathrm{MApr}}_A^{\delta}(X)=\{x\in U\mid \delta_A^*(x)\cap X\neq \varnothing\} \quad (10)$$

在盛魁等人[8]提出的混合型信息系统邻域粗糙集基础上，下面将进一步提出混合信息系统的邻域熵、邻域联合熵、邻域条件熵以及邻域条件互信息熵模型等，进一步完善邻域粗糙集模型下的信息熵理论。

定义6：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ， $A=A_n\cup A_m$ 确定的混合邻域关系为 $\mathrm{MNR}^{\delta}(A)$ ，对象 $x\in U$ 在 $\mathrm{MNR}^{\delta}(A)$ 下的邻域类为 $\delta_A^*(x)$ ，那么混合邻域关系 $\mathrm{MNR}^{\delta}(A)$ 确定的混合邻域熵 $\mathrm{MNE}^{\delta}(A)$ 定义如下。

$$\mathrm{MNE}^{\delta}(A)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|} \quad (11)$$

其中，对象 x_i 的邻域不确定性构成了对象集的邻域熵（即平均不确定性），定义为 $\mathrm{MNE}_{\delta}^{x_i}(A)=-\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}$ 。

定义7：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ，属性子集 $A,B\subseteq C$ ，那么 A 和 B 的混合邻域联合熵定义如下。

$$\mathrm{MNE}^{\delta}(A,B)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|U|} \quad (12)$$

定义8：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ，属性子集 $A,B\subseteq C$ ，那么 B 关于 A 的混合邻域条件熵定义如下。

$$\mathrm{MNE}^{\delta}(B\mid A)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|\delta_A^*(x_i)|} \quad (13)$$

根据定义6~定义8，混合邻域条件熵具有如下性质。

性质1：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ，属性子集 $A,B\subseteq C$ ，那么可以得到式（14）。

$$\mathrm{MNE}^{\delta}(B\mid A)=\mathrm{MNE}^{\delta}(A,B)-\mathrm{MNE}^{\delta}(A) \quad (14)$$

证明：根据定义6和定义7，可以得到式（15）。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A,B)-\mathrm{MNE}^{\delta}(A)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|U|}-\\&\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}\right)=-\frac{1}{|U|}\\&\sum_{i=1}^{|U|}\left(\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|U|}-\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}\right)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|\delta_A^*(x_i)|}=\\&\mathrm{MNE}^{\delta}(B\mid A)\end{aligned} \quad (15)$$

则 $\mathrm{MNE}^{\delta}(B\mid A)=\mathrm{MNE}^{\delta}(A,B)-\mathrm{MNE}^{\delta}(A)$ 成立。

定义8中的混合邻域条件熵与信息论中的条件熵类似，反映了引入属性子集A后B中剩余的不确定性量，混合邻域条件熵可以通过A和B的联合不确定性与A的不确定性来计算。

定义9：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$ ，属性子集 $A,B\subseteq C$ ，那么 A 和 B 的混合邻域互信息熵定义为如下。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A;B)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|\cdot|\delta_B^*(x_i)|}{|U|\cdot|\delta_{A\cup B}^*(x_i)|}\end{aligned} \quad (16)$$

混合邻域熵、混合邻域条件熵和混合邻域互信息熵具有如下关系。

性质2：设混合型信息系统表示为$\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，属性子集$A,B\subseteq C$，那么可以得到如下计算式。

① $\mathrm{MNE}^{\delta}(A;B)=\mathrm{MNE}^{\delta}(B;A)$；

② $\mathrm{MNE}^{\delta}(A;B)=\mathrm{MNE}^{\delta}(A)+\mathrm{MNE}^{\delta}(B)-\mathrm{MNE}^{\delta}(A,B)$；

③$\mathrm{MNE}^{\delta}(A;B)=\mathrm{MNE}^{\delta}(A)-\mathrm{MNE}^{\delta}(A\mid B)=\mathrm{MNE}^{\delta}(B)-\mathrm{MNE}^{\delta}(B\mid A)$。

证明：根据定义9，可以得到式（17）。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A;B)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|\cdot|\delta_B^*(x_i)|}{|U|\cdot|\delta_{A\cup B}^*(x_i)|}=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_B^*(x_i)|\cdot|\delta_A^*(x_i)|}{|U|\cdot|\delta_{B\cup A}^*(x_i)|}=\\&\mathrm{MNE}^{\delta}(B;A)\end{aligned}\tag{17}$$

则①成立。

根据定义6和定义7可以得到：

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A)+\mathrm{MNE}^{\delta}(B)-\mathrm{MNE}^{\delta}(A,B)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_B^*(x_i)|}{|U|}\\&-\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|U|}\right)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|\cdot|\delta_B^*(x_i)|}{|U|\cdot|\delta_{A\cup B}^*(x_i)|}=\\&\mathrm{MNE}^{\delta}(A;B)\end{aligned}\tag{18}$$

则②成立。

根据定义6和定义8可以得到：

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A)-\mathrm{MNE}^{\delta}(A\mid B)\\&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}\\&-\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|\delta_B^*(x_i)|}\right)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\left(\mathrm{lb}\frac{|\delta_A^*(x_i)|}{|U|}-\mathrm{lb}\frac{|\delta_{A\cup B}^*(x_i)|}{|\delta_B^*(x_i)|}\right)=\\&\mathrm{MNE}^{\delta}(A;B)\end{aligned}\tag{19}$$

同理，可以得到：

$$\mathrm{MNE}^{\delta}(A;B)=\mathrm{MNE}^{\delta}(B)-\mathrm{MNE}^{\delta}(B\mid A)\tag{20}$$

则③成立。

通过性质2可以看出属性子集A和B的互信息量与B和A的互信息量是一致的。属性子集A和B混合邻域互信息熵可以表示为各自的混合邻域熵值去除A和B后的混合邻域联合熵值。

与信息论理论类似，接下来进一步提出混合邻域条件互信息熵。

定义10：设混合型信息系统表示为$\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，属性子集$A_1,A_2$，$B\subseteq C$，那么在属性子集$B$下，$A_1$和$A_2$的混合邻域条件互信息熵定义为如下。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A_1;A_2\mid B)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A_1\cup B}^*(x_i)|\cdot|\delta_{A_2\cup B}^*(x_i)|}{|\delta_B^*(x_i)|\cdot|\delta_{A_1\cup A_2\cup B}^*(x_i)|}\end{aligned}\tag{21}$$

混合邻域条件互信息熵具有如下性质。

性质3：设混合型信息系统表示为$\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$,属性子集$A_1,A_2,B\subseteq C$，那么可以得到式（22）。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A_1;A_2\mid B)=\\&\mathrm{MNE}^{\delta}(A_1,B)+\mathrm{MNE}^{\delta}(A_2,B)-\\&\mathrm{MNE}^{\delta}(A_1,A_2,B)-\mathrm{MNE}^{\delta}(B)\end{aligned}\tag{22}$$

证明：根据定义6和定义7，可以得到式（23）。

$$\begin{aligned}&\mathrm{MNE}^{\delta}(A_1,B)+\mathrm{MNE}^{\delta}(A_2,B)-\\&\mathrm{MNE}^{\delta}(A_1,A_2,B)-\mathrm{MNE}^{\delta}(B)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A_1\cup B}^*(x_i)|}{|U|}\\&+\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A_2\cup B}^*(x_i)|}{|U|}\right)\\&-\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A_1\cup A_2\cup B}^*(x_i)|}{|U|}\right)\\&-\left(-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_B^*(x_i)|}{|U|}\right)=\\&-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{A_1\cup B}^*(x_i)|\cdot|\delta_{A_2\cup B}^*(x_i)|}{|\delta_B^*(x_i)|\cdot|\delta_{A_1\cup A_2\cup B}^*(x_i)|}\end{aligned}\tag{23}$$

因此，满足$\mathrm{MNE}^{\delta}(A_1;A_2\mid B)=\mathrm{MNE}^{\delta}(A_1,B)+\mathrm{MNE}^{\delta}(A_2,B)-\mathrm{MNE}^{\delta}(A_1,A_2,B)-\mathrm{MNE}^{\delta}(B)$。

性质3表明，混合邻域条件互信息熵可通过混合邻域熵和混合邻域联合熵计算得到。

性质4：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，属性子集 $A_1,A_2,B\subseteq C$，那么可以得到式（24）。

$$\mathrm{MNE}^{\delta}(A_1;A_2\mid B)=\mathrm{MNE}^{\delta}(A_2;A_1\mid B) \quad (24)$$

证明：根据混合邻域条件互信息熵的定义可以直接得到。

根据性质3可以看出，当属性子集 A_1 和 A_2 相互独立时，混合邻域条件互信息熵的值为0。这表明混合邻域条件互信息熵可以展示给定条件下属性子集之间的依赖程度。将混合邻域条件互信息熵作为信息系统的属性子集评估函数，可以进行混合型信息系统的属性约简。

3 属性约简算法

本节将利用混合邻域条件互信息熵评估信息系统属性之间的依赖度和独立性，并构造出一种混合型信息系统的属性约简算法。

属性约简旨在寻找属性全集中与分类强相关的属性子集，因此属性约简集中的属性与信息系统的类属性具有强相关性。由于互信息熵展示了属性之间的相关性，因此将提出的混合型邻域互信息熵和混合型邻域条件互信息熵用于混合型信息系统属性之间的相关性度量。

定义11：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，属性子集 $A\subseteq C$，关于决策属性集 D 的相关度 $\varphi_D(A)$ 定义如下。

$$\varphi_D(A)=\mathrm{MNE}^{\delta}(A;D) \quad (25)$$

其中，$\mathrm{MNE}^{\delta}(A;D)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_A^*(x_i)|\cdot|[x_i]_D|}{|U|\cdot|\delta_{A\cup D}^*(x_i)|}$，$[x_i]_D$ 为对象 x_i 在决策属性 D 下的等价类。

定义12：设混合型信息系统表示为 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，属性子集 $A\subseteq C$，属性子集 $B\subseteq(C-A)$ 在属性子集 A 下关于决策属性 D 的相关度 $\varphi_D(B|A)$ 定义如下。

$$\varphi_D(B|A)=\mathrm{MNE}^{\delta}(B;A\mid D) \quad (26)$$

其中，$\mathrm{MNE}^{\delta}(B;A\mid D)=-\frac{1}{|U|}\sum_{i=1}^{|U|}\mathrm{lb}\frac{|\delta_{B\cup D}^*(x_i)|\cdot|\delta_{A\cup D}^*(x_i)|}{|[x_i]_D|\cdot|\delta_{B\cup A\cup D}^*(x_i)|}$。

利用混合型邻域互信息熵和混合型邻域条件互信息熵对混合型信息系统进行属性选择，可以进一步设计出一种属性约简算法。

算法1：基于邻域条件互信息熵的混合型信息系统属性约简算法

输入：混合型信息系统 $\mathrm{MIS}=(U,\mathrm{AT}=C\cup D)$，邻域半径 δ。

输出：属性约简结果 red。

1.设置属性约简初始结果 $\mathrm{red}=\varnothing$。

2.对于条件属性集 C 中的每个属性 $\forall a\in C$，计算属性 a 与决策属性集 D 的相关度 $\varphi_D(\{a\})$。

3.找出2中相关度最大的属性 $a_{\max}$，即 $a_{\max}=\arg\max_{a\in C}(\varphi_D(\{a\}))$。

4. 令 $\mathrm{red}\leftarrow\mathrm{red}\cup\{a_{\max}\}$，$C'\leftarrow C-\{a_{\max}\}$。

5.对于属性 $\forall b\in C'$，计算属性 b 在属性约简集 red 下关于决策属性 D 的相关度 $\varphi_D(\{b\}|\mathrm{red})$。

6.找出5中相关度最大的属性 $b_{\max}$，即 $b_{\max}=\arg\max_{b\in C'}(\varphi_D(\{b\}\mid\mathrm{red}))$。

7.令 $\mathrm{red}=\mathrm{red}\cup\{b_{\max}\}$，$C'\leftarrow C'-\{b_{\max}\}$，并利用分类器对属性约简结果 red 进行分类精度计算，记录其分类精度结果。

8.重复5~7，直至 $C'=\varnothing$。

9.找出所有属性约简中分类精度最大的属性约简结果 $\mathrm{red}_{\mathrm{best}}$。

10.返回属性约简集 $\mathrm{red}_{\mathrm{best}}$。

在算法1中，主要计算量集中在属性

集的邻域条件互信息熵上，而邻域条件互信息熵的计算主要是针对对象邻域类的计算，因此整个算法1的时间复杂度为 $O(|\mathrm{AT}|^2 \cdot |U|^2)$。

4 实验分析

为了验证本文提出的基于邻域条件互信息熵的属性约简算法的有效性，下面使用6个数据集进行实验分析，这些数据集见**表1**。这些数据集选择自UCI公共数据集，这些数据集均为混合型类型，适用于本文所提算法。

同时本文选择3种同类型的属性约简算法进行实验，分别为参考文献[6]提出的属性约简算法（对比算法1），参考文献[10]提出的属性约简算法（对比算法2）和参考文献[19]提出的属性约简算法（对比算法3）。

所有算法的属性约简结果通过支持向量机（support vector machine，SVM）分类器和朴素贝叶斯（naive Bayesian，NB）分类器计算其分类精度，对每个数据集的约简结果进行20次十折交叉验证，并将平均值作为最终的分类精度结果。本实验在MATLAB 2018b上对所有属性约简算法进行实现，所有实验都在Intel(R) Core(TM)i3-7100上进行，CPU时钟速率为3.90 GHz，内存为8 GB。

在本文提出的属性约简算法中，不同的邻域半径取值对算法的属性约简结果将产生很大的影响。在参考文献[4-8,19]中，学者们通过大量实验发现，当邻域半径过小时，其属性约简的长度较小，并且分类精度也较小；当邻域半径过大时，其分类精度不会更高。对于数据集归一化为0和1之间的值，当邻域半径为0.15左右时，其属性约简长度不是很大且分类精度最高，因此本实验选择邻域半径为0.15进行后续实验。

表1 实验数据集

名称	样本数/个	属性数/个	分类数目/个
Cylinder	512	40	3
Credit	690	15	2
German	1 000	19	2
Segment	2 310	19	7
Sick	2 800	28	3
Abalone	4 177	8	29

4.1 分类精度结果对比

分类性能是验证属性约简算法质量最有效和最直接的方法，其中，通常利用分类精度来衡量算法分类性能。**表2**和**表3**分别展示了本文属性约简算法与3种对比算法在SVM分类器和NB分类器下的平均分类精度结果，其结果使用“平均值±标准差”的形式表示。

对比**表2**和**表3**的实验结果，可以得到如下结论。

- 与原始数据集的分类精度相比，3种对比算法和本文算法的SVM分类精度分别提高了6%、8%、5%和11%，NB分类精度分别提高了8%、6%、9%和12%。
- 在大部分数据集下，本文的属性约简算法具有更高的分类精度，例如对于利用SVM分类器计算得到的分类精度，本文算法在Cylinder、Credit和Segment等数据集上更高；对于利用NB分类器计算得到的分类精度，本文算法在German、Segment和Sick等数据集上更高。
- 同时本文算法在SVM分类器和NB分类器下的分类精度标准差大多小于或等于其余对比算法。从统计学的角度来看，本文算法的稳定性更高，这主要是由于本文算法通过邻域条件互信息熵选择属性，降低了最终约简结果中的冗余属性，从而提

表 2 SVM 分类精度结果

对比项	原始数据集	对比算法1[6]	对比算法2[10]	对比算法3[19]	本文算法
Cylinder	0.81±0.018	0.86±0.011	0.83±0.024	0.86±0.013	0.89±0.009
Credit	0.72±0.023	0.71±0.008	0.77±0.021	0.76±0.013	0.80±0.013
German	0.73±0.005	0.82±0.014	0.85±0.031	0.86±0.018	0.84±0.016
Segment	0.82±0.009	0.83±0.018	0.86±0.027	0.84±0.017	0.88±0.017
Sick	0.86±0.021	0.93±0.020	0.91±0.021	0.90±0.012	0.95±0.011
Abalone	0.82±0.040	0.89±0.014	0.86±0.043	0.83±0.025	0.89±0.016
平均	0.79±0.019	0.84±0.014	0.85±0.028	0.83±0.016	0.88±0.014

表 3 NB 分类精度结果

对比项	原始数据集	对比算法1[6]	对比算法2[10]	对比算法3[19]	本文算法
Cylinder	0.80±0.021	0.83±0.015	0.82±0.022	0.85±0.011	0.88±0.012
Credit	0.74±0.017	0.77±0.014	0.79±0.026	0.83±0.015	0.81±0.018
German	0.74±0.009	0.86±0.018	0.83±0.027	0.84±0.012	0.87±0.014
Segment	0.80±0.012	0.82±0.015	0.84±0.015	0.83±0.018	0.86±0.016
Sick	0.83±0.014	0.90±0.021	0.88±0.017	0.89±0.014	0.93±0.013
Abalone	0.78±0.032	0.87±0.016	0.84±0.026	0.82±0.019	0.87±0.019
平均	0.78±0.018	0.84±0.016	0.83±0.022	0.85±0.015	0.87±0.015

高了最终约简结果的分类性能。

4.2 属性约简长度对比

对于属性约简算法来说，属性约简结果的长度也是评估算法有效性的一项重要指标，**表4**展示了本文算法与3种对比算法的属性约简长度的对比结果。从**表4**可以看出，本文算法在各个数据集上的平均属性约简集长度为7.7，均低于其余3种算法，说明本文算法能够选择出规模更小的属性约简集。

4.3 属性约简效率对比

此外，算法的效率也是评估算法有效性和实用性的又一重要指标，**图1**给出了各个属性约简算法对每个数据集进行属性约简的用时。由**图1**可以看出，本文算法和对比算法1的用时均小于其余对比算法，这再一次证明了本文算法的有效性和优越性。

4.4 不同邻域半径分类精度结果对比

为了进一步对比本文算法和对比算法在不同邻域半径下属性约简的分类精度结果，下面将邻域半径区间[0.02,0.4]以0.02为间隔，分别取值对各个算法进行属性约简实验，并计算出每个邻域半径属性约简结果的分类精度。**图2~图4**展示出了部分数据集在不同邻域半径下属性约简的分类

表 4 属性约简长度

名称	原始数据集	对比算法1[6]	对比算法2[10]	对比算法3[19]	本文算法
Cylinder	40	12	11	12	10
Credit	15	7	8	8	6
German	19	10	10	12	10
Segment	19	8	7	6	6
Sick	28	13	14	12	10
Abalone	8	4	4	5	4
平均	21.5	9	9	9.2	7.7

精度结果。由图2~图4可以发现，在不同邻域半径下，本文算法的属性约简分类精度整体上高于其余3种对比算法，因此对于不同邻域半径，本文算法仍然具有更高的属性约简性能。

综合各个环节的实验结果，与其他同类型属性约简算法相比，本文提出的属性约简算法具有更显著的有效性和优越性。

5 结束语

针对目前基于粗糙集理论的属性约简

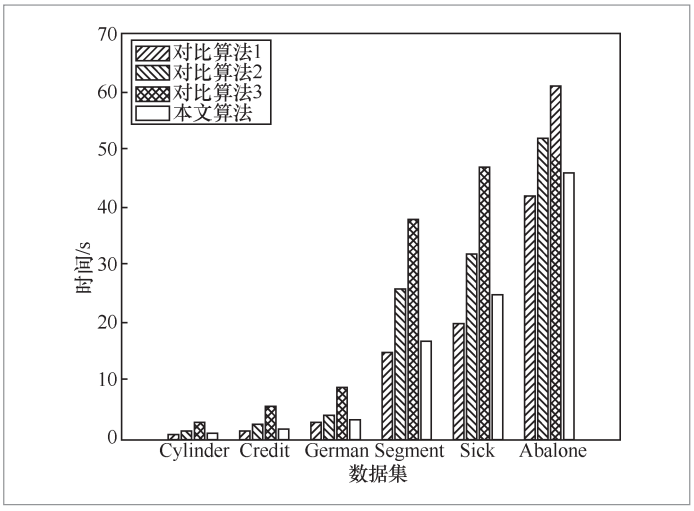

图 1 不同算法运行时间

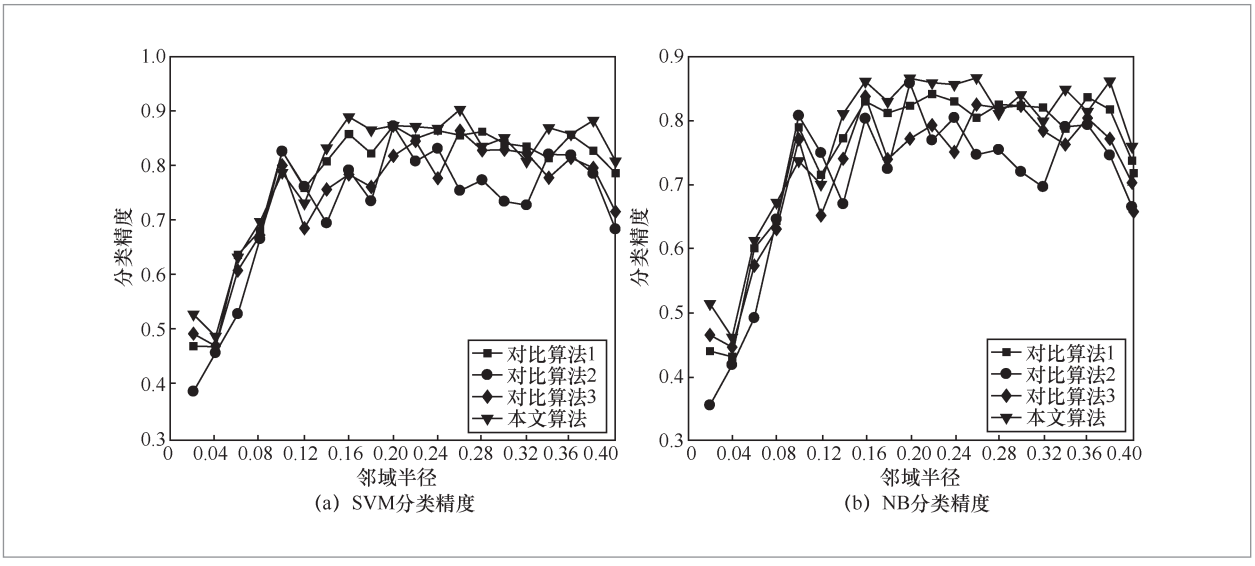

图 2 Cylinder 实验结果

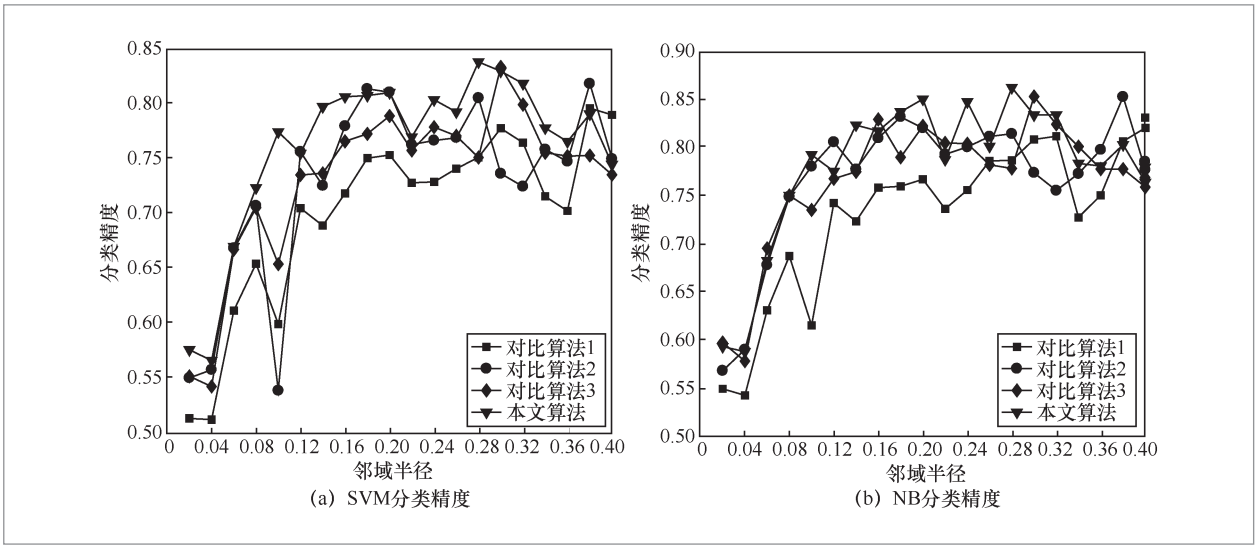

图 3 Credit 实验结果

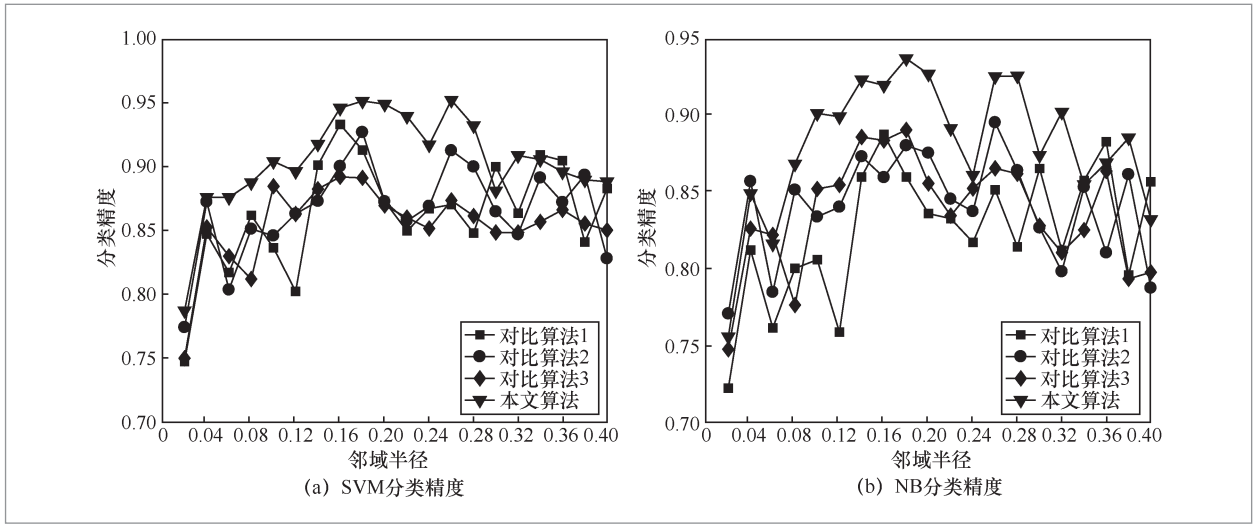

图 4 Sick 实验结果

算法没有考虑属性之间的相关性和依赖性，本文提出一种基于邻域条件互信息熵的混合型信息系统属性约简算法。文中首先在传统邻域熵的基础上进一步提出了混合型邻域互信息熵模型和混合型邻域条件互信息熵模型，然后利用这两种熵模型进行混合型信息系统的属性相关性度量，最后设计出一种新的启发式属性约简算法，基于UCI数据集的属性约简实验表明，所提算法具有更高的属性约简性能。在将来的工作中，笔者将进一步研究邻域互信息熵模型和邻域条件互信息熵模型的增量式属性约简问题。

参考文献:

[1] YANG X, LI M M, FUJITA H, et al. Incremental rough reduction with stable attribute group[J]. Information Sciences, 2022, 589: 283–299.

[2] SUN L, YIN T Y, DING W P, et al. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems[J]. Information Sciences, 2020, 537: 401-424.

[3] 周涛, 陆惠玲, 任海玲, 等. 基于粗糙集的属性约简算法综述[J]. 电子学报, 2021, 49(7): 1439-1449.
ZHOU T, LU H L, REN H L, et al. Survey on attribute reduction algorithm of rough set[J]. Acta Electronica Sinica, 2021, 49(7): 1439-1449.

[4] HU Q H, YU D R, LIU J F, et al. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[5] PANG Q Q, ZHANG L. Semi-supervised neighborhood discrimination index for feature selection[J]. Knowledge-Based Systems, 2020, 204: 106224.

[6] HU M, TSANG E C C, GUO Y T, et al. A novel approach to attribute reduction based on weighted neighborhood rough sets[J]. Knowledge-Based Systems, 2021, 220: 106908.

[7] SHU W H, QIAN W B, XIE Y H. Incremental feature selection for dynamic hybrid data using neighborhood rough set[J]. Knowledge-Based Systems, 2020, 194: 105516.

[8] 盛魁, 王伟, 卞显福, 等. 混合数据的邻域区分度增量式属性约简算法[J]. 电子学报, 2020, 48(4): 682-696.
SHENG K, WANG W, BIAN X F, et al. Neighborhood discernibility degree incremental attribute reduction algorithm for mixed data[J]. Acta Electronica Sinica, 2020, 48(4): 682-696.

[9] 姚晟, 李初宴, 陈悦. 基于非平衡数据下不完备混合型信息系统的属性约简[J]. 计算机应用研究, 2021, 38(5): 1331-1335.
YAO S, LI C Y, CHEN Y. Attribute reduction of incomplete hybrid information system based on unbalanced data[J]. Application Research of Computers, 2021, 38(5): 1331-1335.

[10] WANG C Z, HUANG Y, DING W P, et al. Attribute reduction with fuzzy rough self-information measures[J]. Information Sciences, 2021, 549: 68-86.

[11] YUAN Z, CHEN H M, LI T R, et al. Unsupervised attribute reduction for mixed data based on fuzzy rough sets[J]. Information Sciences, 2021, 572: 67-87.

[12] 栾雨雨, 王锡淮, 肖健梅. 基于混沌离散粒子群的粗糙集属性约简算法[J]. 计算机仿真, 2021, 38(7): 271-275.
LUAN Y Y, WANG X H, XIAO J M. Rough set attribute reduction algorithm based on chaotic discrete particle swarm optimization[J]. Computer Simulation, 2021, 38(7): 271-275.

[13] HU M, TSANG E C C, GUO Y T, et al. Attribute reduction based on overlap degree and k-nearest-neighbor rough sets in decision information systems[J]. Information Sciences, 2022, 584: 301-324.

[14] 桑彬彬, 杨留中, 陈红梅, 等. 优势关系粗糙集增量属性约简算法[J]. 计算机科学, 2020, 47(8): 137-143.
SANG B B, YANG L Z, CHEN H M, et al. Incremental attribute reduction algorithm in dominance-based rough set[J]. Computer Science, 2020, 47(8): 137-143.

[15] 熊菊霞, 吴尽昭, 王秋红. 邻域互信息熵的混合型数据决策代价属性约简[J]. 小型微型计算机系统, 2021, 42(8): 1584-1590.
XIONG J X, WU J Z, WANG Q H. Decision cost attribute reduction of hybrid data based on neighborhood mutual information entropy[J]. Journal of Chinese Computer Systems, 2021, 42(8): 1584-1590.

[16] 陈帅, 张贤勇, 唐玲玉, 等. 邻域互补信息度量及其启发式属性约简[J]. 数据采集与处理, 2020, 35(4): 630-641.
CHEN S, ZHANG X Y, TANG L Y, et al. Neighborhood complementary information measures and heuristic attribute reduction[J]. Journal of Data Acquisition and Processing, 2020, 35(4): 630-641.

[17] 陈帅. 基于三层粒结构的邻域互补信息度量及其属性约简[D]. 成都: 四川师范大学, 2020.
CHEN S. Neighborhood complementary information measures and their attribute reductions based on three-layer granular structure[D]. Chengdu: Sichuan Normal University, 2020.

[18] 姚晟, 徐风, 吴照玉, 等. 基于邻域粗糙互信息熵的非单调性属性约简[J]. 控制与决策, 2019, 34(2): 353-361.
YAO S, XU F, WU Z Y, et al. Non-monotonic attribute reduction based on neighborhood rough mutual information entropy[J]. Control and Decision, 2019, 34(2): 353-361.

[19] SALEM O A M, LIU F, CHEN Y P P, et al. Feature selection and threshold method based on fuzzy joint mutual information[J]. International Journal of Approximate Reasoning, 2021, 132: 107-126

[20] HU Q H, ZHANG L, ZHANG D, et al. Measuring relevance between discrete and continuous features based on neighborhood mutual information[J]. Expert Systems With Applications, 2011, 38(9): 10737-10750.

作者简介

兰海波(1979-), 男, 中国气象局公共气象服务中心高级工程师, 主要研究方向为大数据处理技术、自然语言处理技术、数据库技术和气象服务信息系统的关键技术及应用。

收稿日期: 2022-02-23
通信作者: 兰海波, paper_data001@sina.com