DOI: 10.11992/tis.201804055
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180607.1357.002.html

基于模糊超网络的知识获取方法研究

程麟焰[1,2], 胡峰[1,2]
（1. 重庆邮电大学 计算机科学与技术学院, 重庆 400065; 2. 重庆邮电大学 计算智能重庆市重点实验室, 重庆 400065）

摘　要: 本文结合模糊粗糙集理论与超网络的相关知识, 提出了一种模糊超网络模型。与传统超网络模型的不同之处在于, 模糊超网络模型采用了模糊等效关系来代替超网络中的分明等效关系, 并在此基础上对超边的生成和演化进行了改进。根据样本的分布将样本集划分成 3 个区域, 即正域、边界域和负域, 不同区域的样本按照不同的方式生成超边; 根据分类效果将超边集也划分成 3 个区域, 并对不同区域的超边进行相应地替换处理。实验结果表明, 在正确率、Precision、Recall 等指标上, 模糊超网络分类算法具有明显的优势。
关键词: 模糊等价; 模糊集; 模糊粗糙集; 三支决策; 超网络; 知识获取方法; 分类算法
中图分类号: TP18　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)03−0479−12

中文引用格式: 程麟焰, 胡峰. 基于模糊超网络的知识获取方法研究[J]. 智能系统学报, 2019, 14(3): 479–490.
英文引用格式: CHENG Linyan, HU Feng. Fuzzy hypernetwork-based knowledge acquisition method[J]. CAAI transactions on intelligent systems, 2019, 14(3): 479–490.

Fuzzy hypernetwork-based knowledge acquisition method

CHENG Linyan[1,2], HU Feng[1,2]
(1. College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Combining the fuzzy rough set theory with the related knowledge on hypernetworks, this paper proposes a fuzzy hypernetwork mode. In comparison with the traditional hypernetwork model, the fuzzy hypernetwork model uses the fuzzy equivalence relationship to replace the distinct equivalence relation in hypernetworks and then improves the generation and evolution of hyperedges on this basis. First, the samples are divided into three regions according to their distribution: positive, boundary, and negative regions. The samples of different regions generate hyperedges in different ways. Second, the hyperedges are also divided into three regions according to their classification results, and the corresponding replacement of hyperedges in different regions is implemented. The experimental results show that the fuzzy hypernetwork classification algorithm presents prominent advantages in terms of accuracy, precision, and recall, thus proving the validity of the classification algorithm.
Keywords: fuzzy equivalence; fuzzy set; fuzzy rough set; three-way decision; hypernetworks; knowledge acquisition method; classification algorithm

模糊粗糙集理论是 1990 年由 D.Dubios 和 H.Prade 共同提出的处理数值型数据中存在的不一致性的数学理论[1]。经过多年的发展, 模糊粗糙集在理论和应用方面都取得了相当丰富的研究成果, 在系统控制、故障诊断、机器学习与数据挖掘等众多领域都有着广泛的应用。经典的粗糙集

收稿日期: 2018−04−26.　**网络出版日期:** 2018−06−07.
基金项目: 国家自然科学基金项目 (61533020, 61472056, 61309014); 重点产业共性关键技术创新专项项目 (cstc2017zdcy-zdyf0332, cstc2017zdcy-zdzx0046); 重庆市基础与前沿项目 (cstc2017jcyjAX0408).
通信作者: 程麟焰. E-mail: 496732322@qq.com.

理论强调的是对象间的不可区分性，主要用于处理清晰、离散且有限的属性值，而在实际生活中大部分的数据集都具有多种多样的属性值，粗糙集在处理这些本身具有模糊性的数据和连续属性时存在一定的局限性[2]。粗糙集理论中的等效关系是研究模糊粗糙集理论的基础，将经典粗糙集理论中的被近似对象由清晰集转换为模糊集，并将论域上的分明等效关系弱化为模糊等效关系即可得到模糊粗糙集[3]。

超网络 (hypernetworks) 是受生物分子网络启发而建立的一种基于超图实现的认知学习模型，能够表示模式特征间的高阶关联关系[4]。目前，国内的研究者们主要研究演化超网络模型，探究其应用领域并在此基础上对超网络模型进行改进。王进等[5]结合 OCDD 算法，提出了一种能处理多值数据的细粒度超网络分类方法；王进等[6]在超网络的演化学习过程中引入遗传算法，从而得到了一种具有较高的分类准确率的模式识别方法；为了处理多类型癌症分子问题，王进等[7]提出了一种基于演化超网的多类型癌症分子分型系统。同时，在中文文本分类[8]、评分预测、道路限速标志识别[9]等方面，演化超网络模型也得到了很好的应用。超网络的研究在国内起步较晚，在许多领域都值得研究和学习。

本文结合模糊粗糙集的思想提出了一种模糊超网络 (fuzzy hypernetworks, F-hypernetworks)。在模糊超网络中，对于连续型的属性不需要对其进行离散化处理，解决了传统超网络只能处理离散型属性的问题，并对传统超网络训练过程中具有很大随机性的超边替代环节进行了改进。

1　相关概念

1.1　模糊等价类

定义 1　给定决策表 $(U, A\cup D)$，其中：U 为非空有限论域；A 为条件属性集合，也称特征集合；D 为决策属性集合，也称类别属性集。在没有说明的情况下，属性是指条件属性。$P\subseteq A$ 对应一个不可分辨的等效关系，简记为 P。若 P 满足：

1) 自反性，$\forall x\in U$，$\mu_P(x,x)=1$；

2) 对称性，$\forall x,y\in U$，$\mu_P(x,y)=\mu_P(y,x)$；

则称 P 为 U 上的模糊相似关系[10]。

定义 2　设 P 是 U 上的模糊相似关系，对于给定的 $x\in U$，令 $[x]_P=\mu_P(x,y)$，$y\in U$，则 $[x]_P$ 是论域 U 上的一个模糊集，称其为 x 关于 P 的模糊邻域[10]。$\mu_P(x,y)$ 表示由模糊相似关系 P 确定的对象 x 和 y 之间的模糊相似度，可由式 (1) 确定：

$$\mu_P(x,y)=\left(\sum_{a\in p}\mu_a(x,y)\right)\Big/|P| \tag{1}$$

对于属性 $a\in P$，若 a 为连续属性，则 $\mu_a(x,y)$ 可由式 (2) 表示的模糊相似度确定：

$$\mu_a(x,y)=\exp\left(-\frac{(a(x)-a(y))^2}{2{\sigma_a}^2}\right) \tag{2}$$

式中：σ_a 表示所有对象在属性 a 上取值的标准方差。若 a 为离散属性，则按式 (3)[11]计算模糊相似度：

$$\mu_a(x,y)=\begin{cases}0, & a(x)\neq a(y)\\ 1, & a(x)=a(y)\end{cases} \tag{3}$$

式中 $a(x)$、$a(y)$ 分别表示对象 x、y 在属性 a 上的属性值。

定义 3　给定决策表 $(U, A\cup D)$，P 是 U 上的模糊相似关系，对于给定的 $x\in U$ 有

$$[x]_{P\lambda}=\{y|\mu_P(x,y)\geqslant\lambda,\lambda\in[0,1]\} \tag{4}$$

式中：$[x]_{P\lambda}$ 称为 x 关于 P 的 λ-等价类；实数 λ 为模糊相似度阈值。

定义 4　设 (U,P) 是模糊近似空间，U 为论域，P 是 U 上的模糊相似关系，(U,P) 上的 (I, T)-模糊粗糙近似是一个映射 $\mathrm{Apr}: F(U)\rightarrow F(U)\times F(U)$，任意 $X\in F(U)$，$\mathrm{Apr}: F(X)=(\underline{P_I}X,\overline{P^T}X)$，其中，$\underline{P_I}X$ 称为 X 在 (U,P) 中的 I-下模糊粗糙近似[12]、$\overline{P^T}X$ 称为 T-上模糊粗糙近似[12]，两者的隶属函数描述为

$$\mu_{\underline{P_I}X}(x)=\inf_{y\in U}I(\mu_P(x,y),\mu_X(y)),\ \forall x\in U \tag{5}$$

$$\mu_{\overline{P^T}X}(x)=\sup_{y\in U}T(\mu_P(x,y),\mu_X(y)),\ \forall x\in U \tag{6}$$

对 $\forall y\in U$，若 $y\in X$，则 $\mu_X(y)$ 为 1，否则为 0。$\mu_P(x,y)$ 由式 (1) 确定。其中 I 表示边缘蕴含算子、T 表示 t-模[3]：

$$I(x,y)=\min(1-x+y,1) \tag{7}$$

$$T(x,y)=\max(x+y-1,0) \tag{8}$$

根据式 (5)，对象 x 关于模糊正域的隶属度[13]可表示为

$$\mu_{\mathrm{POS}_P(D)}(x)=\sup_{X\in U/D}\mu_{\underline{P_I}X}(x) \tag{9}$$

在模糊粗糙集条件下，决策属性 D 对条件属性集 P 的依赖度[13]为

$$k=\gamma'_P(D)=\frac{\left|\mu_{\mathrm{POS}_P(D)}(x)\right|}{|U|}=\frac{\sum\limits_{x\in U}\mu_{\mathrm{POS}_P(D)}(x)}{|U|} \tag{10}$$

k 值的大小，反映了条件属性集 P 的分类能力。决策属性 D 对属性集 P 的依赖程度越大，以

P 为依据进行分类的效果越好。以表 1 所示的决策信息系统为例计算各个属性的依赖度。

表 1　决策信息系统

Table 1　Decision information system

样本	a_1	a_2	a_3	D
1	−0.4	−0.3	−0.5	N
2	−0.4	0.2	−0.1	Y
3	−0.3	−0.4	−0.3	N
4	0.3	−0.3	0	Y
5	0.2	−0.3	0	Y
6	0.2	0	0	N

a_1、a_2、a_3 为条件属性, D 为决策属性。对于所有 $x,y\in U$, 根据式 (1) 分别计算关于条件属性 a_1、a_2、a_3 的对象间的模糊相似度:

$$\mu_{a_1}(x,y)=\begin{bmatrix}1.000\,0 & 1.000\,0 & 0.955\,8 & 0.109\,3 & 0.196\,6 & 0.196\,6\\ 1.000\,0 & 1.000\,0 & 0.955\,8 & 0.109\,3 & 0.196\,6 & 0.196\,6\\ 0.955\,8 & 0.955\,8 & 1.000\,0 & 0.196\,6 & 0.323\,2 & 0.323\,2\\ 0.109\,3 & 0.109\,3 & 0.196\,6 & 1.000\,0 & 0.955\,8 & 0.955\,8\\ 0.196\,6 & 0.196\,6 & 0.323\,2 & 0.955\,8 & 1.000\,0 & 1.000\,0\\ 0.196\,6 & 0.196\,6 & 0.323\,2 & 0.955\,8 & 1.000\,0 & 1.000\,0\end{bmatrix}$$

$$\mu_{a_2}(x,y)=\begin{pmatrix}1.000\,0 & 0.097\,4 & 0.911\,0 & 1.000\,0 & 1.000\,0 & 0.432\,4\\ 0.097\,4 & 1.000\,0 & 0.034\,9 & 0.097\,4 & 0.097\,4 & 0.688\,9\\ 0.911\,0 & 0.034\,9 & 1.000\,0 & 0.911\,0 & 0.911\,0 & 0.225\,2\\ 1.000\,0 & 0.097\,4 & 0.911\,0 & 1.000\,0 & 1.000\,0 & 0.432\,4\\ 1.000\,0 & 0.097\,4 & 0.911\,0 & 1.000\,0 & 1.000\,0 & 0.432\,4\\ 0.432\,4 & 0.688\,9 & 0.225\,2 & 0.432\,4 & 0.432\,4 & 1.000\,0\end{pmatrix}$$

$$\mu_{a_3}(x,y)=\begin{pmatrix}1.000\,0 & 0.155\,6 & 0.628\,1 & 0.054\,6 & 0.054\,6 & 0.054\,6\\ 0.155\,6 & 1.000\,0 & 0.628\,1 & 0.890\,2 & 0.890\,2 & 0.890\,2\\ 0.628\,1 & 0.628\,1 & 1.000\,0 & 0.351\,2 & 0.351\,2 & 0.351\,2\\ 0.054\,6 & 0.890\,2 & 0.351\,2 & 1.000\,0 & 1.000\,0 & 1.000\,0\\ 0.054\,6 & 0.890\,2 & 0.351\,2 & 1.000\,0 & 1.000\,0 & 1.000\,0\\ 0.054\,6 & 0.890\,2 & 0.351\,2 & 1.000\,0 & 1.000\,0 & 1.000\,0\end{pmatrix}$$

决策划分:

$$U/D=\{\{1,3,6\},\{2,4,5\}\}=\{X_1,X_2\}$$

$$\mu_{X_1}(x)=\{1,0,1,0,0,1\}$$

$$\mu_{X_2}(x)=\{0,1,0,1,1,0\}$$

根据式 (5) 可得:

$$\mu_{\underline{a_1 X_1}}(1)=0,\ \mu_{\underline{a_1 X_1}}(2)=0,\mu_{\underline{a_1 X_1}}(3)=0.044\,2$$

$$\mu_{\underline{a_1 X_1}}(4)=0,\ \mu_{\underline{a_1 X_1}}(5)=0,\ \mu_{\underline{a_1 X_1}}(6)=0$$

$$\mu_{\underline{a_1 X_1}}(x)=\{0,0,0.044\,2,0,0,0\}$$

同理可得:

$$\mu_{\underline{a_1 X_2}}(x)=\{0,0,0,0.044\,2,0,0\}$$

$$\mu_{\mathrm{POS}_{a_1}(D)}(x)=\sup_{X\in U/D}\mu_{\underline{a_1 X}}(x)=$$

$$\max\{\mu_{\underline{a_1 X_1}}(x),\mu_{\underline{a_1 X_2}}(x)\}=\{0,0,0.044\,2,0.044\,2,0,0\}$$

$$\gamma'_{a_1}(D)=0.014\,7$$

按上述方法分别求出 a_2、a_3 的依赖度:

$$\gamma'_{a_2}(D)=0.118\,5$$

$$\gamma'_{a_3}(D)=0.221\,0$$

由此可以计算出每个属性的依赖度, 并称其为属性的重要度。

1.2　模糊超网络模型

定义 5　设 $G=<X, E, \lambda>$是一个模糊超网络, $X=\{x_1,x_2,\cdots,x_n\}$ 表示模糊超网络的顶点集合, $E=\{e_1,e_2,\cdots,e_m\}$ 为超网络的超边集合, λ 为模糊超网络模型的最优模糊相似度阈值。超边的条件属性集为 $C=\{c_1,c_2,\cdots,c_s\}$, D 为超边的决策属性, e_i 是超边集 E 中连接 k 个顶点 $x_{i1},x_{i2},\cdots,x_{ik}$ 的超边。其中顶点 x_i 为样本, 且一条超边中的样本具有相同的属性集。

定义 6　模糊超网络 $G_1=<X_1, E_1, \lambda_1>$, 模糊超网络 $G_2=<X_2, E_2, \lambda_2>$, 若 $X_1=X_2$ 则 $\lambda_1=\lambda_2$。

定义 7　模糊超网络 $G=<X, E, \lambda>$, 超边的属性集为 $C=\{c_1,c_2,\cdots,c_s\}$, $\forall B(B\subseteq C)$, 在属性集 B 上, 样本 $x=\{c_1(x),c_2(x),c_3(x),\cdots,c_p(x),D(x)\}$, $c_1(x)$, $c_2(x)$, $\cdots$, $c_p(x)$ 表示 x 在属性 c_i 上的取值, $D(x)$ 表示 x 的决策分类。

定义 8　给定模糊超网络 $G=<X, E, \lambda>$, 样本 x 在属性集 $B(B\subseteq C)$ 上的 λ-等价类超边集合为

$$[x]_{B\lambda}=\{e|(e\in E)\wedge\mu_B(x,e)\geqslant\lambda\}\tag{11}$$

λ-等价类样本集合为

$$[x]_B^{\lambda}=\{y|(y\in X)\wedge\mu_B(x,y)\geqslant\lambda\}$$

定义 9　给定模糊超网络 $G=<X, E, \lambda>$, $\forall e\in E$, 在属性集 $B(B\subseteq C)$ 上, 关联超边 e 的样本集合表示为

$$R_{B\lambda}(e)=\{x|e\in[x]_{B\lambda},x\in X\}\tag{12}$$

定义 10　给定模糊超网络 $G=<X, E, \lambda>$, $\forall e\in E$, $D(e)$ 表示超边 e 的决策分类, 在属性集 $B(B\subseteq C)$ 上, 关联超边 e 的样本集合为 $R_{B\lambda}(e)$, 当 $R_{B\lambda}(e)\neq\varnothing$ 时, 超边 e 对样本分类的置信度为

$$\mathrm{Conf}_B=\frac{|\{x|x\in R_{B\lambda}(e),D(x)=D(e)\}|}{|\{x|x\in R_{B\lambda}(e)\}|}\tag{13}$$

定义 11　给定模糊超网络 $G=<X, E, \lambda>$, C 为样本的条件属性集, D 为样本的决策属性, 对任意的样本 $x\in X$ 有:

1) 如果 $f(x)\geqslant\alpha$, 则 $x\in\mathrm{POS}(X)$;

2) 如果 $\beta<f(x)<\alpha$, 则 $x\in\mathrm{BND}(X)$;

3) 如果 $f(x)\leqslant\beta$ 或 $f(x)=-1$, 则 $x\in\mathrm{NEG}(X)$。

$$f(x)=\frac{|\{y|\mu_C(x,y)\geqslant\lambda,D(x)=D(y)\}|}{|\{y|\mu_C(x,y)\geqslant\lambda\}|},\ y\in X\tag{14}$$

如果 $|\{y|\mu_C(x,y)\geqslant\lambda\}|=0$ 则 $f(x)=-1$。$f(x)$ 表示在样本 x 的 λ-等价类样本集合中, 与 x 同类的样

本所占的比例。$f(x)$ 越大，说明 x 的模糊等价类与 x 类别一致的概率越大。

图 1 给出了 4 个样本的 λ-等价类样本集合，由式 (14) 可得：$f(x_1)=1$，$f(x_2)=0.2$，$f(x_3)=0$，$f(x_4)=-1$。本文实验选取 $\alpha=1,\beta=0$，$f(x_1)\geqslant 1$，故 x_1 是正域样本；$0<f(x_2)<1$，x_2 是边界域样本；$f(x_3)\leqslant 0$，x_3 是负域样本；$f(x_4)=-1$，x_4 没有 λ-等价类样本，也是负域样本。

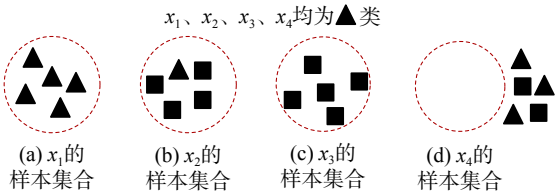

图 1　λ-等价类样本示例

Fig. 1　Examples of λ-equivalence class sample

定义 12　给定模糊超网络 $G=<X, E, \lambda>$，C 为样本的条件属性集，D 为样本的决策属性，任意超边集 $E'(E'\subseteq E)$ 关于属性集 B 的正域、负域和边界域可分别定义为

$$\begin{aligned}&\mathrm{POS}(E')=\{e|\max_{D(x)\neq D(e)}\{\mu_B(x,e)\}<\lambda\cap\\&\qquad\max_{D(x)=D(e)}\{\mu_B(x,e)\}\geqslant\lambda,x\in X,e\in E'\}\\&\mathrm{NEG}(E')=\{e|\max_{D(x)\neq D(e)}\{\mu_B(x,e)\}\geqslant\lambda+\frac{1-\lambda}{2},\\&\qquad x\in X,e\in E'\}\\&\mathrm{BND}(E')=E'-\mathrm{POS}(E')-\mathrm{NEG}(E')\end{aligned}\tag{15}$$

以表 1 的决策信息系统为例，图 2 是表 1 的一个模糊超网络模型，超边集 $E_Y=\{e_1,e_2,e_3,e_4\}$，假设超边与各样本的模糊相似度如表 2 所示，最优模糊相似度阈值为 λ=0.5。图 2 中实线圆区域表示超边的 λ-等价类，虚线圆区域表示该超边的 λ_0-等价类，$\lambda_0=\lambda+(1-\lambda)/2=0.75$。

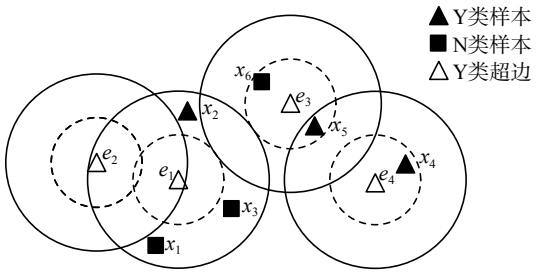

图 2　模糊超网络示例

Fig. 2　Example of a Fuzzy hypergraph

根据式 (15)、表 2 与图 2 可知，正域超边需同时满足两个条件：

1) $\max\limits_{D(x)\neq D(e)}\{\mu_B(x,e)\}<0.5$；

2) $\max\limits_{D(x)=D(e)}\{\mu_B(x,e)\}\geqslant 0.5$。

负域超边需满足条件：

$$\max_{D(x)\neq D(e)}\{\mu_B(x,e)\}\geqslant 0.75$$

表 2　样本-超边相似度

Table 2　Sample_Hyperedge similarity

$\mu(e,x)$	e_1	e_2	e_3	e_4
x_1	0.65	0.30	0.28	0.20
x_2	0.70	0.35	0.46	0.22
x_3	0.72	0.28	0.36	0.34
x_4	0.12	0.10	0.35	0.82
x_5	0.30	0.15	0.77	0.60
x_6	0.35	0.27	0.77	0.33

对于超边 e_1，与 e_1 相似度最高的异类样本为 x_3，$\mu(e_1,x_3)$=0.72<0.75，不满足负域条件，所以 e_1 不是负域超边，$\mu(e_1,x_3)$=0.72>0.5 不满足正域条件 1)，所以 e_1 是边界域超边。

对于超边 e_2，与 e_2 相似度最高的异类样本为 x_1，$\mu(e_2,x_1)$=0.30<0.75，不满足负域条件，所以 e_2 不是负域超边，与 e_2 相似度最高的同类样本为 x_2，$\mu(e_2,x_2)$=0.35<0.5 不满足正域条件 2)，所以 e_2 是边界域超边。

对于超边 e_3，与 e_3 相似度最高的异类样本为 x_6，$\mu(e_3,x_6)$=0.77>0.75，满足负域条件，所以 e_3 是负域超边。

对于超边 e_4，与 e_4 相似度最高的异类样本为 x_3，$\mu(e_4,x_3)$=0.34<0.5，与 e_4 相似度最高的同类样本为 x_4，$\mu(e_4,x_4)$=0.82>0.5 满足正域条件，所以 e_4 是正域超边。

综上所述，$\mathrm{POS}(E_Y)=\{e_4\}$，$\mathrm{BND}(E_Y)=\{e_1,e_2\}$，$\mathrm{NEG}(E_Y)=\{e_3\}$。

2　模糊超网络分类算法

2.1　算法思路

同传统超网络一样，模糊超网络生成算法也分为三大步骤：初始化超边集，训练样本分类，超边替代。超网络通过迭代训练的方式进行演化学习，当分类正确率和迭代次数满足特定条件时，即可退出迭代，输出模型。由于传统超网络采用随机生成的方式初始化超边，增大了超边替代阶段筛选和替换分类能力差的超边的难度[14]。所以本文提出的模糊超网络对超边的初始化随机生成进行了控制，同时在超边替代过程中，对不同域中的超边进行相应的处理以提高超网络的分类效果。算法流程如图 3 所示。

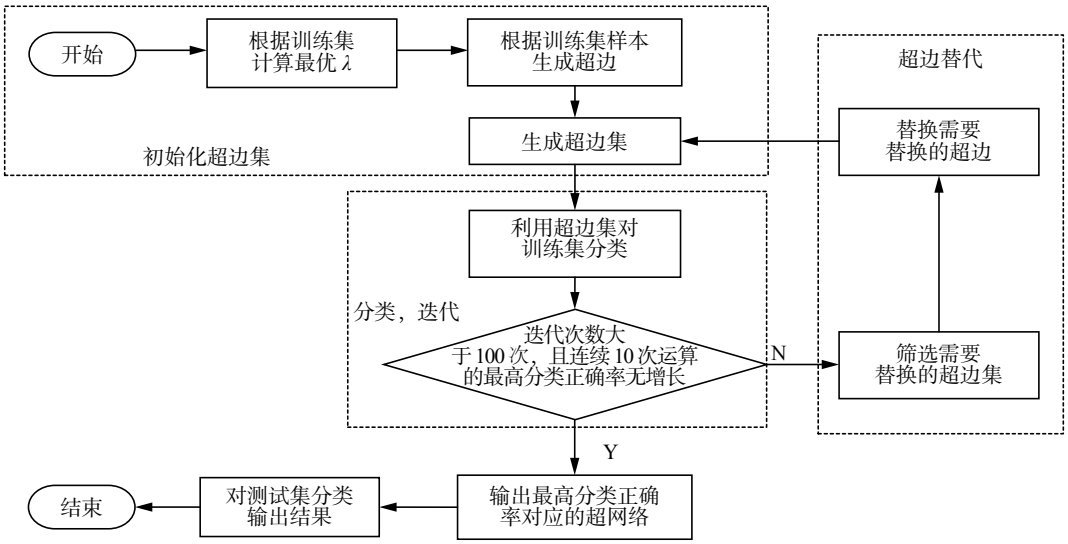

图 3　分类算法流程

Fig. 3　Flow of this algorithm

2.1.1　计算最优模糊相似度阈值 λ

由定义 6 可知, 每一个训练样本集都有且只有一个最优模糊相似度阈值 λ, 所以本文在执行分类算法前需要通过循环迭代的方法计算出最优 λ, 具体流程如图 4 所示。初始设置模糊相似度阈值 λ_0 为 0, 然后通过叠加步长来改变 λ_0 的取值, 在不同的 λ_0 值下, 采用模糊超网络分类方法对训练集进行十折交叉验证[15]得到相应的分类正确率。将正确率最高的 λ_0 值作为最优模糊相似度阈值 λ 执行后续的分类算法。值得注意之处在于, 从理论上说, 对于同一个训练集, λ 是唯一的, 本方法计算出的结果仅是一个接近的阈值, 一般步长设置越短越接近最优模糊相似度阈值。本文所设置的步长 s=0.01, 足以满足实验需求。

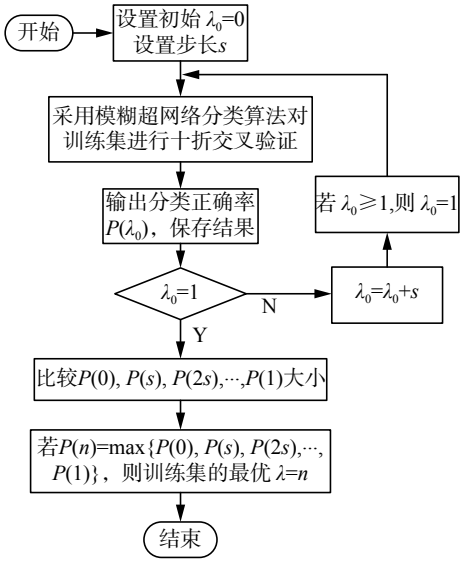

图 4　计算最优 λ 流程图

Fig. 4　Flow chart for calculating optimal λ

2.1.2　超边初始化

根据训练集中的样本生成模糊超网络中的超边。本文设置每个样本直接生成 5 条超边, 超边的属性数目与样本一致。每条超边的初始化主要由条件属性初始化和决策属性初始化两部分组成。

1) 条件属性初始化

条件属性初始化主要有两种方式: 一种是随机属性继承, 超边从条件属性集中随机选择十分之七的属性继承样本的属性值, 即超边在这些属性上的取值与生成该超边的样本相同。剩余属性则根据训练集在该属性上的取值范围随机生成属性值。如图 5 所示, x 为样本, e 为 x 按照随机属性继承方式生成的超边。

x	1	2	3	4	5	6	7	8	9	10	
e	1	2	10	66	5	6	7	8	9	12	

图 5　随机属性继承示例图

Fig. 5　Example of random attribute inheritance

另一种是择优属性继承, 超边从所有属性中选择重要度较高的前十分之七的属性继承样本的属性值, 剩余属性上的取值则根据训练集中同类样本在该属性上的取值范围随机生成。如图 6 所示, 样本 x 拥有 10 个属性, 首先利用样本 x 的 λ-等价类样本集合按照定义 4 所示的方法计算出各个属性的重要度 k, 然后重新生成重要度较低的属性 1、2、9 对应的属性值。

k	0.35	0.23	0.55	1.00	0.54	0.44	0.78	0.40	0.28	0.66	
x	1	2	3	4	5	6	7	8	9	10	
e	33	55	3	4	5	6	7	8	12	10	

图 6　择优属性继承示例图

Fig. 6　Example of preferred attribute inheritance

2) 决策属性初始化

正域样本生成的超边，条件属性初始化采取随机属性继承方式，决策属性直接继承生成该超边的样本。边界域样本生成的超边，条件属性初始化采取择优属性继承方式，决策属性直接继承生成该超边的样本。负域样本生成的超边，同样采取随机属性继承的方式确定条件属性，因为其决策属性与原始样本相同的概率较低，所以该类超边的决策属性是根据整个数据集确定的，与关联该超边的样本集中的大类样本类别一致。

2.1.3 训练样本分类

给定模糊超网络 $G=<X, E, \lambda>$，决策属性的取值范围 $V_D=\{1,2,\cdots,m\}$ 样本 x 在 λ 下的模糊等价类超边集合为 $[x]_\lambda$，其中决策属性为 j 的超边集合为

$$[x]_\lambda^j=\{e|e\in[x]_\lambda\cap D(e)=j\} \tag{16}$$

对于每一个样本 x 的分类过程如下：

1) 计算出样本 x 在 λ 下的模糊等价类超边 $[x]_\lambda$；

2) 将 $[x]_\lambda$ 中的超边进行分类，将决策属性为 j 的超边归类到 $[x]_\lambda^j$ 中，

$$|[x]_\lambda|=\sum_{j\in V_D}\left|[x]_\lambda^j\right| \tag{17}$$

3) 计算 x 的类别 $D(x)$：

$$D(x)=\operatorname{argmax}\left(\left|[x]_\lambda^j\right|\right),\ j\in V_D$$

4) 若 $[x]_\lambda=\varnothing$，则选取与样本 x 模糊相似度最高的 n 条超边 $e_1,e_2,\cdots,e_n$ 加入到集合 $E_n(x)$ 中，类别 $D(x)=\operatorname{argmax}\left(\left|E_n(x)^j\right|\right), j\in V_D$，本文实验取 n=3。

采用上述分类规则对训练集的每个样本进行分类，计算模糊超网络模型对训练集的分类正确率。

2.1.4 超边替代

首先判断超边所在区域，然后不同的区域采取不同的措施：

若超边 e 是正域超边，则超边 e 与训练集中所有异类样本相似度较低，与同类样本相似度较高，该超边的存在有助于提高分类效果，故选择保留超边 e；

若超边 e 是负域超边，则超边 e 与训练集中某一异类样本相似度较高，在分类的过程中超边 e 会影响其他类别的分类效果，故需要替换；

若超边 e 是边界域超边，在衡量超边 e 的分类效果时，需要通过置信度 Conf_B 进行判断，若 $\mathrm{Conf}_B>\gamma$(本文取 γ=0.5)，保留超边 e，反之，则替换掉超边 e。其中存在一种特殊情况：当关联超边 e 的样本数为 0 时，无法计算置信度 Conf_B。此时，需要查看超边参与分类的样本集合 $P(e)$，$P(e)=\{x|[x]_\lambda=\varnothing\wedge e\in E_n(x), x\in X\}$，若 $|P(e)|=0$ 说明超边 e 在超网络对训练集的分类过程中没有起太大的作用，一般选择将其替换；若 $|P(e)|\neq 0$，则按式 (18) 计算置信度：

$$\mathrm{Conf}_B=\frac{\{x|D(e)=D(x), x\in P(e)\}}{|P(e)|} \tag{18}$$

2.2 算法描述

算法 1 初始化超边库算法

输入 训练样本集 X，最优阈值 λ；

输出 超边集 E。

1) $E=\varnothing$

2) 计算每个样本的 λ-等价类样本集合，根据定义 11 判断样本所属区域

3) for each x in X do

4) j=0

5) while(j<5) /*设置每个样本生成的超边数*/

6) if $x\in$ POS(X) then

7) 根据样本 x 生成超边 e，e 直接继承 x 的决策属性，条件属性初始化采取随机属性继承方式

8) end if

9) if $x\in$ NEG(X) then

10) 根据样本 x 生成超边 e，条件属性初始化采取随机属性继承方式，决策属性根据整个数据集确定

11) end if

12) if $x\in$ BND(X) then

13) 根据 x 生成超边 e，超边直接继承 x 的决策属性，计算在 x 的 λ-等价类样本集合中各条件属性的依赖度，将条件属性按依赖度大小从大到小排序，超边 e 择优选择排在前 7/10 的属性继承样本 x 的属性值

14) end if

15) $E=E\cup\{e\}$; j++

16) end while

17) end for

18) return E; /*输出初始超边库*/

算法 2 超边替代算法

输入 训练样本集 X，最优阈值 λ，超边集 E；

输出 超边集 E。

1) for each e in E do

2) 根据定义 12 判断超边 e 所属区域

3) if $e\in$ POS(E) then

4) end if

5) if $e\in$ BND(E) then

6) 计算 e 的置信度 Conf_B

7) if $\mathrm{Conf}_B \leqslant 0.5$ then

8) 参照算法 1 中的生成超边方法, 该超边对应的原始样本重新生成一条新超边 e_{new}, $E=E-\{e\}$; $E=E\cup\{e_{new}\}$

9) end if

10) end if

11) if $e \in \mathrm{NEG}(E)$ then

12) 参照算法 1, 原始样本重新生成一条新超边 e_{new}, $E=E-\{e\}$; $E=E\cup\{e_{new}\}$

13) end if

14) end for

15) return E; /*输出更新后的超边库*/

算法 3　生成模糊超网络算法

输入　训练样本集 X, 最优阈值 λ;

输出　超网络 G。

1) 执行算法 1 生成初始超边集 E

2) k=0, E_{best}=∅, P_{max}=0 /*用于保存最高分类正确率*/

3) while (k<100)

/*设置最小迭代次数用以保障超网络能够得到充分的演化, 可根据实际演化效果增减次数*/

4) for each x in X do

5) 计算样本 x 在 λ 下的模糊等价类超边集合 $[x]_\lambda$, 计算 x 的类别 $D(x)$

6) end for

7) 根据训练样本的实际类别, 计算当前超边集对训练集的分类正确率 P; k++

8) if $P>P_{max}$ then

9) $P_{max}=P$, $E_{best}=E$; /*保存当前最优超边集*/

10) end if

11) 执行算法 2 更新超边集

12) end while

13) m=0

14) while (m<10) /*退出迭代条件 */

15) for each x in X do

16) 计算样本 x 在 λ 下的模糊等价类超边集合 $[x]_\lambda$, 计算 x 的类别 $D(x)$

17) end for

18) 根据训练样本的实际类别, 计算当前超边集对训练集的分类正确率 P

19) if $P > P_{max}$ then

20) $P_{max}=P$, $E_{best}=E$, m=0

21) else m++

22) end if

23) 执行算法 2 更新超边集

24) end while

26) return $G=<X, E_{best}, \lambda>$; /*输出模糊超网络*/

令训练集样本数目为 n, 样本的属性数目为 m。算法 1 的平均时间复杂度为 $O(m\times n^2)$, 算法 2 的时间复杂度为 $O(m\times n)$, 算法 3 的时间复杂度为 $O(m\times n^2)$, 所以建模的时间复杂度为 $O(m\times n^2)$。

计算最优模糊相似度阈值时循环迭代所用的时间约为建模时间的 $10\times s^{-1}$ 倍, 迭代步长 s 在 (0, 1) 范围内取值。s 值越小, 计算所用的时间越长, 得到的结果越接近理论上的最优阈值。最优模糊相似度阈值是一个非常重要的参数, 在部分数据集上, 略微改变阈值, 分类结果就会有较大的改变。对于这类数据集而言, 计算最优模糊相似度阈值是非常重要的环节。但也有一些数据集, 例如大部分的离散型数据集, 在一定范围内改变阈值, 生成的超网络的分类效果不变, 处理这些数据集时, 可以通过设置合理的步长和初始阈值, 达到既不影响分类效果又能减少迭代时间的目的。

3　实验评价

3.1　数据集及评价指标

为验证算法的有效性, 本文选取机器学习数据库 UCI 中的 15 个数据集进行实验, 每个数据集的详细信息如表 3 所示, 按数据集大小从小到大排序。

表 3　实验数据集

Table 3　Experimental data sets

序号	数据集	属性	类别数	样本数
1	BLOGGER	5N	2	100
2	lymph	3C 15N	4	148
3	tae	1C 4N	3	151
4	flags	2C 26N	8	194
5	Glass	9C	7	214
6	breast-cancer	9N	2	286
7	Haberman	3C	2	306
8	column_2C_weka	6C	2	310
9	column_3C_weka	6C	3	310
10	ecoli	7C	8	336
11	Ionosphere	34C	2	351
12	balance-scale	4N	3	625
13	Pima_diabetes	8C	2	768
14	tic-tac-toe	9N	2	958
15	car	6N	4	1 728

注: C 为 continuous, N 为 nominal

混合矩阵 (confusion matrix)[16]是一个常用的评价指标, 如表 4 所示。TP 表示正类样本被正确预测为正类的样本数; FN、FP 分别表示预测错误的实际正类和负类样本数目; TN 表示负类样本被正确预测为负类的样本数。

表 4　混合矩阵

Table 4　Confusion matrix

类别	预测为正类	预测为负类
正类	TP	FN
负类	FP	TN

查准率: 表示被分类器预测为正类的样本中正类样本所占的比例。计算公式为

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

查全率: 又称召回率, 表示正类样本被分类器正确预测为正类的比例。计算公式为

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

F-Measure 指标[17]是一种综合查全率和查准率的分类评价指标:

$$\text{F}-\text{Measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

正确率即分类正确样本数与分类总数之比。

本文采用正确率、查准率 (Precision)、查全率 (Recall)、和 F-Measure 作为评价指标。

3.2　实验方法

为了考察模糊超网络分类算法的性能, 本文采用 Java 语言实现算法并将其与 NaiveBayes、KNN、J48(C4.5) 、SMO[18] 和 BP-KNN[19] 5 种算法进行对比。利用 Weka[20]平台, 在 15 个数据集上对以上算法进行对比实验, BP-KNN 根据文献[19]设置参数, 其余分类器的参数均为 Weka 平台下算法的默认值。实验中模糊超网络模型所涉及的随机种子均设为 seed=1 234。所有实验结果均为采用 5-折交叉验证后的结果。

3.3　实验结果

本次实验将模糊超网络分类算法封装成 Weka 平台可识别的分类器, 所有的评价指标均由 Weka 平台的评估器计算并输出。表 5~8 分别给出了本算法与其他算法的正确率、查准率、查全率和 F-Measure 的结果。

表 5　正确率值

Table 5　Accuracy　　%

序号	数据集	NaiveBayes	KNN	J48	SMO	BP-KNN	F-hypernetworks
1	BLOGGER	72.000 0	82.000 0	73.000 0	73.000 0	73.000 0	85.000 0
2	lymph	81.756 8	83.108 1	73.648 6	85.810 8	83.783 8	82.432 4
3	tae	52.317 9	65.562 9	51.655 6	56.953 6	48.344 4	56.291 4
4	flags	56.185 6	56.701 0	57.216 5	55.670 1	58.762 9	59.278 4
5	Glass	47.196 3	68.224 3	66.822 4	56.074 8	63.551 4	67.289 7
6	breast-cancer	73.076 9	70.629 4	69.230 8	70.279 7	73.426 6	73.426 6
7	Haberman	74.836 6	67.973 9	73.529 4	73.529 4	70.588 2	72.875 8
8	column_2C_weka	78.064 5	78.709 7	80.322 6	80.000 0	80.322 6	80.967 7
9	column_3C_weka	82.903 2	78.387 1	83.225 8	75.483 9	75.806 5	81.935 5
10	ecoli	85.119 0	80.654 8	81.250 0	82.440 5	85.714 3	86.011 9
11	Ionosphere	82.336 2	87.179 5	90.598 3	88.034 2	86.609 7	88.319 1
12	balance-scale	91.680 0	82.720 0	66.720 0	91.200 0	82.880 0	91.040 0
13	Pima_diabetes	75.390 6	71.484 4	73.046 9	76.562 5	73.046 9	74.869 8
14	tic-tac-toe	70.146 1	98.434 2	84.342 4	98.329 9	98.434 2	84.655 5
15	car	85.300 9	92.245 4	90.740 7	93.344 9	92.245 4	90.509 3
16	平均值	73.887 4	77.601 0	74.356 7	77.114 3	76.434 5	78.326 9

表 6　查准率

Table 6　Precision

序号	数据集	NaiveBayes	KNN	J48	SMO	BP-KNN	F-hypernetworks
1	BLOGGER	0.702	0.817	0.716	0.715	0.717	0.848
2	lymph	0.816	0.833	0.735	0.860	0.830	0.826
3	tae	0.518	0.657	0.518	0.580	0.487	0.568
4	flags	0.582	0.564	0.562	0.546	0.539	0.557
5	Glass	0.482	0.680	0.661	0.510	0.613	0.661
6	breast-cancer	0.716	0.681	0.662	0.677	0.714	0.712
7	Haberman	0.715	0.664	0.688	0.541	0.624	0.656
8	column_2C_weka	0.826	0.801	0.804	0.797	0.804	0.818
9	column_3C_weka	0.826	0.789	0.831	0.772	0.759	0.822
10	ecoli	0.850	0.802	0.801	0.806	0.848	0.850
11	Ionosphere	0.840	0.882	0.907	0.886	0.877	0.893
12	balance-scale	0.845	0.765	0.616	0.919	0.766	0.839
13	Pima_diabetes	0.749	0.711	0.727	0.760	0.722	0.747
14	tic-tac-toe	0.687	0.985	0.842	0.984	0.985	0.845
15	car	0.846	0.921	0.907	0.935	0.921	0.904
16	平均值	0.733	0.770	0.732	0.753	0.747	0.770

表 7　查全率

Table 7　Recall

序号	数据集	NaiveBayes	KNN	J48	SMO	BP-KNN	F-hypernetworks
1	BLOGGER	0.720	0.820	0.730	0.730	0.730	0.850
2	lymph	0.818	0.831	0.736	0.858	0.838	0.824
3	tae	0.523	0.656	0.517	0.570	0.483	0.563
4	flags	0.562	0.567	0.572	0.557	0.588	0.593
5	Glass	0.472	0.682	0.668	0.561	0.636	0.673
6	breast-cancer	0.731	0.706	0.692	0.703	0.734	0.734
7	Haberman	0.748	0.680	0.735	0.735	0.706	0.729
8	column_2C_weka	0.781	0.787	0.803	0.800	0.803	0.810
9	column_3C_weka	0.829	0.784	0.832	0.755	0.758	0.819
10	ecoli	0.851	0.807	0.813	0.824	0.857	0.860
11	Ionosphere	0.823	0.872	0.906	0.880	0.866	0.883
12	balance-scale	0.917	0.827	0.667	0.912	0.829	0.910
13	Pima_diabetes	0.754	0.715	0.730	0.766	0.730	0.749
14	tic-tac-toe	0.701	0.984	0.843	0.983	0.984	0.847
15	car	0.853	0.922	0.907	0.933	0.922	0.905
16	平均值	0.739	0.776	0.743	0.771	0.764	0.783

表 8　F-Measure 值

Table 8　F-Measure

序号	数据集	NaiveBayes	KNN	J48	SMO	BP-KNN	F-hypernetworks
1	BLOGGER	0.699	0.812	0.716	0.702	0.696	0.848
2	lymph	0.812	0.830	0.733	0.857	0.827	0.823
3	tae	0.519	0.656	0.510	0.572	0.481	0.564
4	flags	0.558	0.564	0.560	0.550	0.551	0.561
5	Glass	0.440	0.680	0.663	0.513	0.616	0.665
6	breast-cancer	0.720	0.685	0.668	0.682	0.691	0.701
7	Haberman	0.703	0.671	0.678	0.623	0.641	0.646
8	column_2C_weka	0.788	0.791	0.803	0.788	0.803	0.812
9	column_3C_weka	0.826	0.786	0.832	0.727	0.758	0.820
10	ecoli	0.849	0.803	0.806	0.798	0.851	0.853
11	Ionosphere	0.826	0.866	0.905	0.876	0.860	0.878
12	balance-scale	0.879	0.794	0.640	0.881	0.795	0.873
13	Pima_diabetes	0.751	0.713	0.729	0.756	0.723	0.728
14	tic-tac-toe	0.684	0.984	0.841	0.983	0.984	0.843
15	car	0.843	0.916	0.907	0.934	0.916	0.899
16	平均值	0.726	0.770	0.733	0.749	0.746	0.768

由图 7 可知，相较于其他 5 种分类算法，本文提出的模糊超网络算法在正确率、查准率、查全率上都具有一定的优势。为了进一步考查模糊超网络分类算法与对比算法之间的比较效果，本文根据这 6 种算法在各个评价指标上的实验结果，按照 1、2、3、4、5、6 的次序进行 Rank 排序评分，得到了如表 9 所示的 6 种算法在 15 个实验数据集上的各项 Rank 平均值。

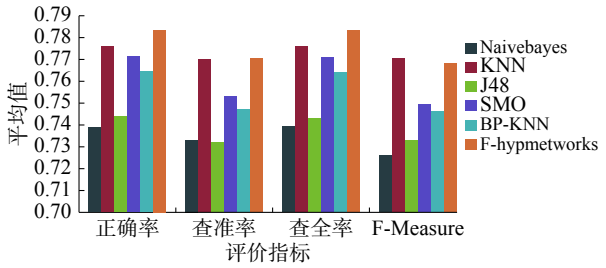

图 7　各项指标平均值柱状图

Fig. 7　Average value of each indicator

表 9　不同算法在评价指标上的 Rank 平均值

Table 9　Rank mean of different algorithms in evaluation index

算法	正确率	查准率	查全率	F-Measure
NaiveBayes	4.133 3	3.333 3	4.133 3	3.866 7
KNN	3.600 0	3.133 3	3.600 0	3.000 0
J48	3.733 3	3.800 0	3.733 3	3.600 0
SMO	3.200 0	3.533 3	3.200 0	3.666 7
BP-KNN	3.200 0	3.866 7	3.200 0	3.866 7
F-hypergraph	2.533 3	2.933 3	2.533 3	2.733 3

从表 5~8 可以看出，本算法在一些数据集上的分类效果并不是最好的。例如，在 tic-tac-toe 数据集上，与分类效果较好的 3 个算法相比，F-hypernetworks 的正确率下降了近 13%，查准率、查全率和 F-Measure 上的结果也相差较大。通过分析可知，在 tic-tac-toe 数据集中，不同类别的样本之间的相似度较高，约 83% 的样本与异类样本的最大模糊相似度不低于其与同类样本的最大模糊相似度。而在分类效果较好的 BLOGGER 数据集上，这种样本的数目只占总数的 22%，在 breast-cancer 数据集上仅有 2%。由于超边对这些样本的识别率不高，模糊超网络对其做出误判的概率较大，如果数据集中存在大量的这种样本，会导致最终的分类结果不理想。

结合图 7 与表 9 可知，模糊超网络在正确率、查准率、查全率、F-Measure 的 Rank 平均值上均为最优。本文选取的对比算法都是应用较为广泛的分类算法，对实际生活中的大部分数据集具有较好分类效果。同这些优秀的算法相比，模糊超网络分类算法仍具有一定的优势，在 Recall、Precision 等指标上表现出了比较好的结果。

4　结束语

本文结合模糊粗糙集和超网络的相关知识提出了一种模糊超网络模型，用于处理分类问题。首先，根据模型的最优模糊相似度阈值 λ 计算样本的 λ-等价类样本集合；其次，根据该集合的分布情况来定义边界域样本、正域样本和负域样本，对于不同区域的样本采取不同的处理方法生成超

边; 然后, 在超边替代阶段, 同样通过划分 3 个区域来控制超边的替换; 最后, 在分类时, 会根据样本的 λ-等价类超边来判断样本的类别。通过在 15 个 UCI 数据集上的实验结果表明, 模糊超网络具有较高的适用性, 在不同的数据集上都能取得较好的分类效果。但是, 随着训练样本数量的增大, 模糊超网络模型在初始化阶段生成的超边越多, 模型在演化训练阶段所消耗的时间越长, 最终会导致算法的运行时间较长, 因此提高算法处理大数据的时间效率将是接下来的研究重点。

参考文献:

[1] RIZA L S, JANUSZ A, BERGMEIR C, et al. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package “RoughSets”[J]. Information sciences, 2014, 287: 68–89.

[2] ZHANG Yu. Research on extension of the fuzzy rough set theory[J]. Advanced materials research, 2012, 532-533: 1472–1476.

[3] 陈德刚. 模糊粗糙集理论与方法[M]. 北京: 科学出版社, 2013.

[4] 王进, 朱文晓, 孙开伟, 等. 基于残差超网络的 DNA 微阵列数据分类[J]. 重庆邮电大学学报 (自然科学版), 2015, 27(5): 647–653.
WANG Jin, ZHU Wenxiao, SUN Kaiwei, et al. Using residual hypernetwork for the classification of DNA microarray data[J]. Journal of Chongqing University of Posts and Telecommunications (natural science edition), 2015, 27(5): 647–653.

[5] 王进, 张军, 胡白帆. 结合最优类别信息离散的细粒度超网络微阵列数据分类[J]. 上海交通大学学报, 2013, 47(12): 1856–1862.
WANG Jin, ZHANG Jun, HU Baifan. Optimal class-dependent discretization-based fine-grain hypernetworks for classification of microarray data[J]. Journal of Shanghai Jiaotong University, 2013, 47(12): 1856–1862.

[6] 王进, 黄萍丽, 孙开伟, 等. 基于演化学习超网络的微阵列数据分类[J]. 江苏大学学报 (自然科学版), 2014, 35(1): 56–62.
WANG Jin, HUANG Pingli, SUN Kaiwei, et al. Microarray data classification based on evolutionary learning hypernetwork[J]. Journal of Jiangsu University (natural science edition), 2014, 35(1): 56–62.

[7] 王进, 丁凌, 孙开伟, 等. 演化超网络在多类型癌症分子分型中的应用[J]. 电子与信息学报, 2013, 35(10): 2425–2431.
WANG Jin, DING Ling, SUN Kaiwei, et al. Applying evolutionary hypernetworks for multiclass molecular classification of cancer[J]. Journal of electronics and information technology, 2013, 35(10): 2425–2431.

[8] 王进, 金理雄, 孙开伟. 基于演化超网络的中文文本分类方法[J]. 江苏大学学报 (自然科学版), 2013, 34(2): 196–201.
WANG Jin, JIN Lixiong, SUN Kaiwei. Chinese text categorization based on evolutionary hypernetwork[J]. Journal of Jiangsu University (natural science edition), 2013, 34(2): 196–201.

[9] 王进, 孙开伟, 李钟浩. 超网络道路限速标志识别[J]. 小型微型计算机系统, 2012, 33(12): 2709–2714.
WANG Jin, SUN Kaiwei, LI Zhonghao. Hypernetworks for road speed limit sign recognition[J]. Journal of Chinese computer systems, 2012, 33(12): 2709–2714.

[10] 齐亚丽. 基于模糊粗糙集属性约简方法的研究[D]. 锦州: 渤海大学, 2016.
QI Yali. The research of attribute reduction method based on fuzzy rough sets[D]. Jinzhou: Bohai University, 2016.

[11] LI Xingyi, LI Xueling, SHI Huaji. Case based reasoning based on fuzzy rough set[C]//Proceedings of the 2nd IEEE International Conference on Information and Financial Engineering. Chongqing, China, 2010: 778–782.

[12] 王世强, 张登福, 毕笃彦, 等. 基于模糊粗糙集和蜂群算法的属性约简[J]. 中南大学学报 (自然科学版), 2013, 44(1): 172–178.
WANG Shiqiang, ZHANG Dengfu, BI Duyan, et al. Attribute reduction method based on fuzzy rough sets and artificial bee colony algorithm[J]. Journal of Central South University (science and technology), 2013, 44(1): 172–178.

[13] WANG Xueen, HAN Deqiang, HAN Chongzhao. Fuzzy-rough set based attribute reduction with a simple fuzzification method[C]//Proceedings of the 24th Chinese Control and Decision Conference. Taiyuan, China, 2012: 3793–3797.

[14] HU Feng, SHI Jin. Neighborhood hypergraph based classification algorithm for incomplete information system [J/OL]. Mathematical problems in engineering, 2015, Article ID 735014, DOI: 10. 1155/2015/735014.

[15] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA, 1995: 1137–1143.

[16] HU Feng, LIU Xiao, LU Xi. A novel cost sensitive classification algorithm based on neighborhood hypergraph[J]. Journal of computational information systems, 2015, 11(1): 109–121.

[17] HU Feng, LI Hang. A novel boundary oversampling al-

gorithm based on neighborhood rough set model: NRS-Boundary-SMOTE[J]. Mathematical problems in engineering, 2013, 2013: 694809.

[18] LUO Yueguo, XIONG Zhongyang, XIA Shuyin, et al. Classification noise detection based SMO algorithm[J]. Optik-international journal for light and electron optics, 2016, 127(17): 7021–7029.

[19] 路敦利, 宁芊, 臧军. 基于 BP 神经网络决策的 KNN 改进算法[J]. 计算机应用, 2017(S2): 65–67.
LU Dunli, NING Qian, ZANG Jun. Improved KNN algorithm based on BP neural network decision making[J]. Journal of computer applications, 2017(S2): 65–67.

[20] 袁梅宇. 数据挖掘与机器学习: WEKA 应用技术与实践[M]. 2 版. 北京: 清华大学出版社, 2014.

作者简介：

程麟焰, 女, 1993 年生, 硕士研究生, 主要研究方向为机器学习与数据挖掘。

胡峰, 男, 1978 年生, 教授, 博士, 主要研究方向为数据挖掘、Rough 集和粒计算。发表学术论文 40 余篇, 被 SCI、EI 检索 20 余篇。

2019 第二届算法、计算和人工智能国际会议 (ACAI 2019)
2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2019)

ACAI 2019 将于 2019 年 12 月 20 日至 22 日在中国三亚召开, 本会议主要围绕“算法、计算和人工智能”的最新研究领域而展开, 致力于促进世界顶尖创新者、科学家、学者、研究人员和思想领导者之间的交流和探讨, 促进数据科学与信息技术领域的发展, 在会议的这三天里, 您将有机会聆听到前沿的学术报告, 见证该领域的成果与进步。

ACAI 2018 已于 2018 年成功在中国三亚召开, 会议论文集由 ACM 出版, 并已成功被 EI 和 Scopus 检索 (http://www.acai2019.net/history.html)。

出版与检索：

所有被接收的文章将被收录在 ACAI 2019 会议论文集中, 并提交 Ei Compendex、SCOPUS、DBLP and Semantic Scholar 等数据库检索。优秀论文可推荐至国际期刊上发表。

征文投递方式：

1. 通过 CMT 投稿系统提交 PDF 版本：https://cmt3.research.microsoft.com/ACAI2019

2. 直接将您的文章或摘要投到我们的会议邮箱, 我们收到后会第一时间回复您 。

投稿邮箱：acai@aiase.net

接受/拒稿通知：论文投稿后 1 ~ 2 周。

联系方式：

会议官网：http://www.acai2019.net/

邮箱：acai@aiase.net

QQ 咨询：3268368942

电话：+852 53465620