文章编号：1000-8608(2014)02-0240-06

基于模糊概念相似性与模糊熵度量的模糊分类算法

冯兴华[1]， 刘晓东*[1]， 刘亚清[2]

（1.大连理工大学 控制科学与工程学院，辽宁 大连 116024；
2.大连海事大学 信息科学技术学院，辽宁 大连 116026）

摘要： 在AFS(axiomatic fuzzy set)理论框架下，提出了一种基于模糊概念相似性与模糊熵度量的分类算法.模糊分类规则的前件通过概念聚合得到，一种基于模糊概念相似性与模糊熵度量的概念选择函数指导聚合过程；然后，利用剪枝算法对得到的模糊规则集进行剪枝，得到最终的分类规则集.用8组来自UCI数据库的数据集作为实验数据对算法进行验证，并与7种经典分类方法进行比较.实验结果表明该算法能得到较高的分类精度，分类结果明显优于参照的分类方法.

关键词： 分类；模糊规则；相似性；模糊熵；公理化模糊集

中图分类号： TP39 **文献标识码：** A **doi：** 10.7511/dllgxb201402014

0 引 言

利用IF-THEN规则的分类方法的优势在于给出分类结果的同时还能够给出分类的语义解释[1].为了克服数据中存在的不确定性和含糊性等，基于模糊规则的分类方法被提出，并被大量研究[2].例如，Wang等[3]提出了基于关联规则的模糊分类器，而Alcala等[4]提出了针对高维数据的模糊关联规则分类方法；Huhn等[5]提出了名为FURIA的分类方法，此方法先产生经典分类规则集，然后用梯形模糊集对分类规则进行模糊化，最后用得到的模糊规则集对数据进行分类.最近，一种由支持向量机诱导分类规则的方法[6]被提出，该方法主要考虑了支持向量机的分类边界及训练样本的分布.很多利用模糊规则分类的方法都是基于模糊概念的.然而，恰当地确定模糊概念的隶属函数是非常困难的，即使是领域内专家也很难做到[7].采用不同的隶属函数对最终分类结果影响很大；而用不同的隶属函数，相同的分类算法也会给出不同的分类结果.这在一定程度上影响了模糊分类器的设计.

基于上述分析，本文在AFS(axiomatic fuzzy set)理论框架下提出一种基于模糊概念相似性与模糊熵度量的分类算法；并在8组来自UCI[8]数据库的实验数据上和其他7种分类算法进行比较，以证明该算法的可靠性和优越性.

1 AFS理论相关知识

设X是一个数据集，即$X=\{x_1,x_2,\cdots,x_N\}$.其中$x_i=\{x_{i,1},x_{i,2},\cdots,x_{i,n}\}$，$x_{i,j}$表示样本$x_i$在第$j$个属性上的取值，$i=1,2,\cdots,N$，$j=1,2,\cdots,n$.在每个属性上，取定若干个简单的模糊概念形成模糊概念集合M.

记$EM^*=\left\{\sum_{i\in I}\left(\prod_{m\in A_i}m\right)\middle|A_i\subseteq M,I\text{是非空指标集}\right\}$.$\prod$表示$A$中元素的合取运算，$\sum$表示$\prod_{m\in A}m$之间的析取运算.文献[9]在集合$EM^*$上定义了一个等价关系$R$：$\sum_{i\in I}\left(\prod_{m\in A_i}m\right),\sum_{j\in J}\left(\prod_{m\in B_j}m\right)\in EM^*$，$\left(\sum_{i\in I}\left(\prod_{m\in A_i}m\right),\sum_{j\in J}\left(\prod_{m\in B_j}m\right)\right)\in R\Leftrightarrow\forall i\in I,\exists k\in J$使得$A_i\supseteq B_k$同时$\forall j\in J,\exists q\in I$使得$B_j\supseteq A_q$.把商集$EM^*/R$记为$EM$.

文献[9]证明了如果如下定义EM上的运算

收稿日期： 2013-06-07； **修回日期：** 2014-01-12.
基金项目： 国家自然科学基金资助项目(61175041).
作者简介： 冯兴华(1982-)，男，博士，E-mail：feng_xinghua@163.com；刘晓东*(1963-)，男，教授，博士生导师，E-mail：xdliuros@dlut.edu.cn.

$\wedge$ 和 $\vee$，则$(EM, \wedge, \vee)$是完全分配格：任意

$$\sum_{i\in I}\left(\prod_{m\in A_i} m\right), \sum_{j\in J}\left(\prod_{m\in B_j} m\right),$$

$$\sum_{i\in I}\left(\prod_{m\in A_i} m\right)\vee\sum_{j\in J}\left(\prod_{m\in B_j} m\right)=\sum_{k\in U}\left(\prod_{m\in C_k} m\right) \quad (1)$$

$$\sum_{i\in I}\left(\prod_{m\in A_i} m\right)\wedge\sum_{j\in J}\left(\prod_{m\in B_j} m\right)=\sum_{i\in I, j\in J}\left(\prod_{m\in A_i\cup B_j} m\right) \quad (2)$$

其中 U 是指标集 I 与 J 的不交并.

设 X 为数据集，M 是 X 上的概念集，对于任意的概念集合 $A\subseteq M$ 及 $x\in X$，定义 X 的子集 $A^{\geqslant}(x)$：

$$A^{\geqslant}(x)=\{y\in X \mid x\geqslant_A y\}\subseteq X \quad (3)$$

式中：$x\geqslant_A y$ 表示对于任意的模糊概念 $m\in A$，x 隶属于 m 的程度大于或者等于 y 隶属于 m 的程度；$A^{\geqslant}(x)$ 完全由概念集合 A 中的概念的语义及样本集合 X 的分布决定.

定义 1[9-10]　设集合 M 是一个概念集合，$(EM,\wedge,\vee)$是完全分配格，则对于任意的 $x\in X$，模糊概念 $\xi=\sum_{i\in I}\left(\prod_{m\in A} m\right)\in EM$ 的隶属函数定义为

$$\mu_\xi(x)=\sup_{i\in I}\frac{|A_i^{\geqslant}(x)|}{|X|} \quad (4)$$

上面定义的模糊概念的隶属函数完全由概念的语义及样本数据的分布决定. 因为格$(EM, \wedge, \vee)$对逻辑运算 $\wedge$、$\vee$ 封闭，所以对于 EM 中的任何概念的隶属函数都可以由上述定义得到.

2　基于模糊规则的分类算法

首先描述指导概念聚合过程的模糊概念选择函数，然后给出产生模糊 IF-THEN 规则集的聚合算法，最后介绍模糊规则集的剪枝算法及利用模糊规则集对样本的分类过程.

2.1　选择函数

(1)模糊概念的相似性

设 α、β 是两个模糊概念，Jaccard 相似性[11] 在 AFS 理论框架下表示为

$$S(\alpha,\beta)=\sum_{x\in X}\mu_{\alpha\wedge\beta}(x)\Big/\sum_{x\in X}\mu_{\alpha\vee\beta}(x) \quad (5)$$

其中 $\vee$、$\wedge$ 分别由式(1) 和(2) 给出. $\mu_{\alpha\wedge\beta}(x)$ 和 $\mu_{\alpha\vee\beta}(x)$ 分别表示样本 x 隶属于模糊概念 $\alpha\wedge\beta$ 和 $\alpha\vee\beta$ 的隶属度，由式(4) 给出.

(2) 模糊熵

α 是一个模糊概念，Λ 是一个模糊概念集合，$\Lambda=\{\beta_1,\beta_2,\cdots,\beta_k\}$. Λ 对 α 的一个划分记为 $\Lambda\mid\alpha$，定义为 $\Lambda\mid\alpha=\{\beta_1\wedge\alpha,\beta_2\wedge\alpha,\cdots,\beta_k\wedge\alpha\}$. 信息熵是衡量一个集合的混乱程度的常用度量. 为了衡量划分 $\Lambda\mid\alpha$ 中 $\beta_i\wedge\alpha$ 与其他部分的相异程度，定义 $\beta_i\wedge\alpha$ 与其他部分之间的模糊熵为

$$Entropy(\alpha,\beta_i)=-\left(\frac{M(\beta_i\wedge\alpha)}{M(\alpha)}\log\frac{M(\beta_i\wedge\alpha)}{M(\alpha)}+\frac{\sum_{j\neq i}M(\beta_j\wedge\alpha)}{M(\alpha)}\log\frac{\sum_{j\neq i}M(\beta_j\wedge\alpha)}{M(\alpha)}\right)$$

这里 $M(\alpha)=\sum_{x\in X}\mu_\alpha(x)$. 事实上，用模糊概念集合 Λ 对 α 进行划分时，$Entropy(\alpha,\beta_i)$ 衡量 $\Lambda\mid\alpha$ 中 $\beta_i\wedge\alpha$ 和其他部分是否是可区分的. $\beta_i\wedge\alpha$ 与其他部分的可区分度越高，$Entropy(\alpha,\beta_i)$ 的取值越小.

(3) 选择函数

如下定义选择函数 $f(\alpha,\beta_i)$：

$$f(\alpha,\beta_i)=S(\alpha,\beta_i)-Entropy(\alpha,\beta_i) \quad (6)$$

式(6) 所示的选择函数评价模糊概念 α 对模糊概念 β_i 的描述能力. 一方面考虑了 α 与 β_i 的相似性，另一方面考虑在用 Λ 对 α 划分时，$\beta_i\wedge\alpha$ 与其他部分的相异性. 一般而言，α 与 β_i 越相似，并且模糊划分 $\Lambda\mid\alpha$ 中 $\beta_i\wedge\alpha$ 与其他部分的相异性越大，选择函数 $f(\alpha,\beta_i)$ 的值越大.

在分类问题中，$\Lambda=\{\beta_1,\beta_2,\cdots,\beta_k\}$ 可以表示决策属性上的类别概念集，β_i 表示第 i 类. 这种情况下，式(6) 就可以用来衡量模糊概念 α 对类别 β_i 的描述能力. 而基于规则的分类方法就是要找到条件属性上的概念 α 来描述决策属性上的类别概念 β_i. 也就是说，α 不但要和 β_i 相似，又要使 $\beta_i\wedge\alpha$ 和 $\beta_j\wedge\alpha(j\neq i)$ 较好地区分开.

下面用图 1 所示的例子来例证所定义的选择函数的合理性. 图 1 所示的是一个两类别分类问题，包含 100 个样本，两类的决策概念分别用 β_1、β_2 表示，即决策属性上的决策概念集为 $\Lambda=\{\beta_1,\beta_2\}$. α_1、α_2、α_3、α_4 是 4 个条件属性上的模糊概念. 表 1 给出了 α_1、α_2、α_3、α_4 分别对 β_1 的选择函数值.

在图 1 中，与 α_3 相比，第一类的 50 个样本在 α_2 上的隶属度与 β_1 更为相似，但是 α_2 对第一类样本与第二类样本的区分度并不大. 从与 β_1 的相似性与对两类样本的区分程度两方面来考虑的话，α_3 对 β_1 的描述能力更强. 也就是说，α_3 比 α_2 更适合作为 β_1 的规则前件. 从表 1 中的数据可以看到 $f(\alpha_3,\beta_1)=-0.10$，$f(\alpha_2,\beta_1)=-0.22$，这说明本文定义的选择函数给出的选择结论与本文的直观分析完全一致.

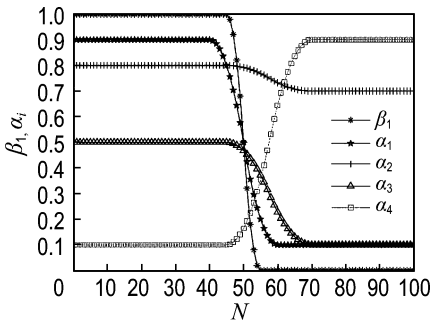

图 1 模糊概念的隶属度函数

Fig. 1 The membership functions of fuzzy concepts

表 1 选择函数在不同模糊概念上的取值

Tab. 1 The selection index values on different fuzzy concepts

α_i	$S(\alpha_i, \beta_1)$	$Entropy(\alpha_i, \beta_1)$	$f(\alpha_i, \beta_1)$
α_1	0.80	0.37	0.43
α_2	0.47	0.69	−0.22
α_3	0.45	0.55	−0.10
α_4	0.06	0.36	−0.30

比较 α_1 与 α_4 的隶属度函数可以看到，虽然 α_1 与 α_4 都可以将两类样本很好地区分开，但是 α_4 与目标类别 β_1 相似性非常小. 这说明与 α_4 相比，α_1 能更好描述 β_1.

从表 1 给出的各个模糊概念对 β_1 的选择函数值可以看到，$f(\alpha_1, \beta_1) > f(\alpha_3, \beta_1) > f(\alpha_2, \beta_1) > f(\alpha_4, \beta_1)$. 这与从图 1 中得到的直观分析结果一致.

2.2 分类算法

本文提出的分类算法利用 IF-THEN 规则对样本进行分类，其中每一条分类规则都具有如下形式：IF α, THEN β. α 表示条件属性上的模糊概念，β 表示决策属性上的概念(模糊或非模糊). 每一条 IF-THEN 规则都是前件和后件之间的关系的体现.

一般情况下，如果分类规则集的前件过于简单，往往不能得到较好的分类精度；如果过于复杂，则意味着有可能发生了过度训练. 本文通过简单概念的聚合得到恰当的规则前件. 通过概念聚合[12]，简单概念被聚合为复杂概念，以便更好地描述规则后件，从而提高分类精度.

概念聚合包含 3 个部分：目标概念、候选概念，以及逻辑算子. 模糊概念的聚合是通过概念的逻辑运算完成的. 如前文所述，在 AFS 理论中定义了两种模糊概念的逻辑算子，分别是“or”$\vee$(如式(1)所示)和“and”$\wedge$(如式(2)所示). 通过逻辑算子将目标概念与候选概念聚合为新的模糊概念，聚合得到的模糊概念的隶属度由式(4)给出.

分类问题涉及的数据可以规范为如下形式：设 $X = \{x_1, x_2, \cdots, x_N\}$ 表示样本集合，$A^1, A^2, \cdots, A^n$ 表示样本集上的 n 个条件属性，C 表示决策属性. 在条件属性 A^i 上取 k_i 个简单模糊概念，用 $T(A^i) = \{m_1^i, m_2^i, \cdots, m_{k_i}^i\}$ 表示 A^i 上的简单概念集合，$T(C) = \{\beta_1, \beta_2, \cdots, \beta_c\}$ 表示决策属性上的决策概念集合，即 $T(C)$ 中的每一个概念表示样本集的一个类别.

模糊分类规则构建算法为每一个类别通过聚合找到最合适的描述概念. 以第一类为例的算法如图 2 所示. 由图 2 给出的算法可以得到一个规则集，其中的每一条规则的前件是输出 *Result* 中的一个模糊概念，后件是第一类的决策概念 β_1.

在图 2 中，Ξ 表示目标概念集；Θ 表示待候选概念集；Ω 表示简单概念集；Ψ 表示算子集；$\Theta \cup \Omega$ 表示候选概念集.

算法的时间复杂度为 $O(N \mid \Omega_0 \mid)$，其中 N 为训练样本的个数，$\mid \Omega_0 \mid$ 为初始简单概念的个数. 算法的最大计算量来自于聚合过程(见图 2)，在每次循环过程中，用 L 个目标概念和至多 $\mid \Omega_0 \mid + T$ 个候选概念进行聚合，聚合过程循环 *length* 次. 而 L、T 和 *length* 三个参数由用户给出，是固定的数值，在下一节的实验中分别设置为 3、2 和 5. 所以时间复杂度为 $O(N \mid \Omega_0 \mid)$.

2.3 推理过程

从图 2 所示算法直接得到的模糊规则集可能包含冗余和分类能力差的规则. 因此，需要将这样的规则从规则集中剔除以提高规则集的分类准确率和分类效率. 这里采用一种剪枝方法来修剪规则集，这种方法只需要规则集和训练样本集的参与. 具体过程如下：

步骤 1 对规则集中的每一条规则进行如下过程：删除该规则，用剩余规则对训练样本进行分类并记录分类准确率及该条规则.

步骤 2 找到步骤 1 中得到的准确率中的最高准确率对应的规则，真正将之删除.

步骤 3 重复步骤 1 及步骤 2 直到如果要真正删除某条规则，规则集对训练样本的分类准确率将出现下降为止.

剪枝后的规则集就是最终用于分类的规则集. 当需要对新样本 x 进行分类时，用式(4)计算该样本在每一条规则的前件上的隶属度. 样本 x 将被分配到规则 r 所表示的类，如果 x 在规则 r 的前件上取得最大的隶属度值.

$Result$ = miningrules(Θ,Ξ,Ω,Ψ,L,T,$length$)

Input:　Ξ 表示目标概念集;Θ 表示待候选概念集;Ω 表示条件属性上的简单概念集;Ψ 表示算子集;
L 表示目标概念集中模糊概念的个数;
T 表示待候选概念集中模糊概念的个数;
$length$ 表示规则前件的最大长度;

Output:　$Result$ 表示聚合得到的规则前件;

// 初始化过程

$\Psi=\{\wedge,\vee\}$;

$\Xi_0=\{\varnothing\}$;

$\Theta_0=\{\}$;

$\Omega_0=\{m_1^1,m_2^1,\cdots,m_{k_1}^1,\cdots,m_1^n,m_2^n,\cdots,m_{k_n}^n\}$;

// 聚合过程

for $i=1$: $length$

$\{\{\zeta_{i1},\varphi_{i1},\eta_{i1}\},\cdots,\{\zeta_{iL},\varphi_{iL},\eta_{iL}\}\}=\operatorname{argmax} L_{\zeta\in\Xi_{i-1},\varphi\in\Psi,\eta\in\Theta_{i-1}\cup\Omega_{i-1},\eta\notin\Xi_{i-1}} f(\zeta\varphi\eta,\beta_1)$;

$\Xi_i=\{\zeta_{i1}\varphi_{i1}\eta_{i1},\cdots,\zeta_{iL}\varphi_{iL}\eta_{iL}\}$;　// 更新结果集

$\{\{\zeta_{i1},\varphi_{i1},\eta_{i1}\},\cdots,\{\zeta_{iT},\varphi_{iT},\eta_{iT}\}\}=\operatorname{argmax} T_{\zeta\in\Xi_{i-1},\varphi\in\Psi,\eta\in\Theta_{i-1}\cup\Omega_{i-1},\eta\notin\Xi_{i-1}} f(\zeta\varphi\eta,\beta_1)$;

$\Theta_i=\{\zeta_{i1}\varphi_{i1}\eta_{i1},\cdots,\zeta_{iT}\varphi_{iT}\eta_{iT}\}$;　// 更新待候选概念集

$\Omega_i=\{m\mid m\in\Omega_{i-1},m\notin\Theta_{i-1}\}$;　// 更新简单概念集

if $\max\{f(\alpha,\beta_1)\mid\alpha\in\Xi_i\}\leqslant\max\{f(\alpha,\beta_1)\mid\alpha\in\Xi_{i-1}\}$　// 停止条件

$i=i-1$;

break;

end

end

$Result=\Xi_i$;

return

图 2　模糊规则构建算法

Fig. 2　The fuzzy rule mining algorithm

3　实验结果及分析

用 8 组来自 UCI 的数据对提出的分类算法进行验证. 表 2 给出了这 8 组数据的基本信息. 在这 8 组数据上,对本文分类方法和其他 7 种分类方法做了比较. 这 7 种方法是 C4. 5[13]、Decision Table (DTable)[14]、Fast Effective Rule Induction (JRip)[15]、Nearest-neighbor-like Algorithm (NNge)[16]、OneR[17]、PART Decision List (PART)[18], 以及 Ripple-down Rule Learner (Ridor)[19]. 上述 7 种分类方法由分类工具箱 Weka[20] 实现且各个分类器均采用工具箱指定的默认参数. 在本文提出的分类方法中,3 个参数分别为 $L=3$,$T=2$,$length=5$. 另外,为简单起见在每个数据集的各个条件属性上取 3 个简单模糊概念,每一个简单模糊概念都有一个明确的含义:离第 i 个分割点较近. 第一个分割点取为属性的最小值;第二个分割点取为属性均值;第三个分割点取为属性的最大值,所有概念的隶属函数由式(4)给出.

在实验中,对每一个分类数据均采用 10 折交叉验证法来评估分类算法分类准确率. 实验得到的分类准确率如表 3 所示. 由表 3 可以看出,本文提出的分类算法(在表 3 中用 AFS 表示)能给出最好的分类结果,8 组数据上的平均准确率最高. 为了进一步分析本文提出的算法与其他 7 种分类算法之间的差异,采用统计分析的方法对这 8 种分类算法得到的结果做出分析.

表 2　来自 UCI 数据库的 8 组数据的基本信息

Tab. 2　The descriptions of datasets from UCI repository

No.	数据集	类别数	样本数	属性数
1	Balance	3	625	4
2	Haberman	2	306	3
3	Iris	3	150	4
4	Liver	2	345	6
5	Wine	3	178	12
6	Teaching	3	151	5
7	Sonar	2	208	60
8	Transfusion	2	748	5

首先,用 Friedman 检验法[21] 来检验这 8 种分类器在 8 组实验数据上得到的分类结果是否存在显著差异;如果这些分类结果存在显著差异,进一步采用 Holm 检验法[21] 来检验本文提出的算法是否和其他方法存在显著差异.

表 3 不同分类方法的比较分析

Tab. 3 A comparative analysis of different classification methods

	分类准确率/%							
	C4.5	DTable	JRip	NNge	OneR	PART	Ridor	AFS
Balance	77.10(6)	74.07(7)	80.64(3)	78.55(5)	57.11(8)	82.55(2)	78.70(4)	**84.66**(1)
Haberman	73.86(3)	73.19(5)	72.18(6)	66.30(8)	**75.11**(1)	68.35(7)	73.49(4)	74.48(2)
Iris	94.00(4)	90.00(8)	92.00(7)	**96.00**(1.5)	92.67(6)	93.33(5)	94.67(3)	**96.00**(1.5)
Liver	62.01(4)	57.39(8)	66.06(2)	59.13(6)	58.53(7)	61.44(5)	**66.88**(1)	64.05(3)
Wine	90.52(5.5)	86.05(7)	92.75(3)	**95.49**(1.5)	78.10(8)	91.63(4)	90.52(5.5)	**95.49**(1.5)
Teaching	52.29(2)	41.67(7)	36.38(8)	**63.58**(1)	42.29(6)	48.96(3)	45.00(5)	47.62(4)
Sonar	74.48(3)	73.93(5)	**78.36**(1)	69.14(7)	55.76(8)	74.00(4)	75.45(2)	73.12(6)
Transfusion	78.48(2)	75.14(7)	**79.15**(1)	70.59(8)	76.35(4)	77.95(3)	75.94(6)	76.21(5)
均值	75.34(3.69)	71.43(6.75)	74.69(3.88)	74.85(4.75)	66.99(6.00)	74.78(4.13)	75.08(3.81)	**76.45**(3.00)

Friedman 检验法的零假设为多个分类方法得到的分类结果不存在显著差异. 在每一个数据集上,按照各个分类方法得到的分类准确率将其排序,每一个分类方法将得到一个序数值(如果分类准确率一样,则均分序数值). 以下假设有 k 个分类方法,实验数据有 N 组. 用 r_i^j 表示第 j 种分类方法在第 i 个数据上的序数值,则 $R_j = (1/N)\sum_i r_i^j$ 表示第 j 种分类方法在 N 组实验数据上的平均序数值. Friedman 检验的统计量为

$$\chi_F^2 = \frac{12}{k(k+1)}\left(\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right)$$

该统计量服从自由度为 $k-1$ 的卡方分布. 如果 k 和 N 比较小,则使用下面的统计量:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{7}$$

统计量 F_F 服从自由度为 $(k-1)$ 和 $(k-1)\cdot(N-1)$ 的 F 分布.

本文所做的实验中 $k=8, N=8, F_F=2.57$,在显著性水平 $\alpha=0.05$ 下统计量的值超过了临界值. 所以,可以拒绝零假设,即这 8 种方法在 8 组数据上得到的分类结果存在显著差异. 然后,用 Holm 测试来分析本文方法和其他 7 种方法之间是否存在显著差异. 分析的结果显示,在显著水平 $\alpha=0.1$ 时,本文方法在 8 组实验数据上的分类结果显著好于 DTable 方法和 OneR 方法. 虽然本文的方法不是显著好于其他 5 种方法,但是表 3 列出的序数值说明,本文的方法取得了最小的均序值,即本文提出的分类算法能得到最好的分类准确率.

4 结 语

本文提出了一种产生模糊分类规则的方法. 该方法通过简单模糊概念的聚合产生模糊分类规则的前件. 一个基于模糊概念相似性与模糊熵的选择函数指导概念的聚合过程. 用 AFS 理论来处理模糊概念的逻辑运算,并给出其隶属函数. 在 AFS 理论框架下,模糊概念的隶属函数完全由概念的语义和样本数据的分布决定,不需要背景知识的参与.

在 8 组来自 UCI 数据库的数据上和 7 种经典分类方法做了比较. 实验结果显示,本文提出的分类算法能得到最好的分类精度,其分类准确率显著地高于 DTable 方法和 OneR 方法.

在以后的工作中,将研究概念聚合的过程,设计新的算法以提高规则集的分类效果. 另一个可以扩展的工作是对规则集的结构的分析,这也可帮助改进聚合算法.

参考文献:

[1] Alsabti K, Ranka S, Singh V. CLOUDS: A decision tree classifier for large datasets [C] // **Conference of Knowledge Discovery and Data Mining.** New York: AAAI Press, 1998: 2-8.

[2] Cano J, Herrera F, Lozano M. Evolutionary stratified training set selection for extracting classification rules with trade off precision interpretability [J]. **Data and Knowledge Engineering**, 2007, **60**(1): 90-108.

[3] WANG Xin, LIU Xiao-dong, Pedrycz W. Mining axiomatic fuzzy set association rules for classification problems [J]. **European Journal of Operational Research**, 2012, **218**(1): 202-210.

[4] Alcala F J, Alcala R, Herrera F. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning [J]. **IEEE Transactions on Fuzzy Systems**, 2011, **19**(5): 857-872.

[5] Huhn J, Hullermeier E. FURIA: An algorithm for unordered fuzzy rule induction [J]. **Data Mining and Knowledge Discovery**, 2009, **19**(3): 293-319.

[6] ZHU Peng-fei, HU Qing-hua. Rule extraction from support vector machines based on consistent region covering reduction [J]. **Knowledge-Based Systems**, 2013, **42**(1): 1-8.

[7] Chandra B, Varghese P P. Fuzzy SLIQ decision tree algorithm [J]. **IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics**, 2008, **38**(5): 1294-1301.

[8] Merz C J, Murphy P M. UCI repository for machine learning data-bases [Z]. Irvine: Department of Information and Computer Science, University of California, 1996.

[9] LIU Xiao-dong. The fuzzy theory based on AFS algebras and AFS structure [J]. **Journal of Mathematical Analysis and Applications**, 1998, **217**(2): 459-478.

[10] LIU Xiao-dong, WANG Wei, CHAI Tian-you. The fuzzy clustering analysis based on AFS theory [J]. **IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics**, 2005, **35**(5): 1013-1027.

[11] Tan P N, Steinbach M, Kumar V. **Introduction to Data Mining** [M]. MA: Addison-Wesley, 2005.

[12] HUANG Zhi-heng, Gedeon T D, Nikravesh M. Pattern trees induction: A new machine learning method [J]. **IEEE Transactions on Fuzzy Systems**, 2008, **16**(4): 958-970.

[13] Quinlan J R. **C4.5: Programs for Machine Learning** [M]. San Mateo: Morgan Kaufmann, 1993.

[14] Kohavi R. The power of decision tables [C] // **European Conference on Machine Learning**. Heraclion: Springer, 1995: 174-189.

[15] Cohen W W. Fast effective rule induction [C] // **The Twelfth International Conference on Machine Learning**. California: Morgan Kaufmann, 1995: 115-123.

[16] Roy S. **Nearest Neighbor with Generalization** [M]. Christchurch: University of Canterbury, 2002.

[17] Holte R C. Very simple classification rules perform well on most commonly used datasets [J]. **Machine Learning**, 1993, **11**(1): 63-91.

[18] Frank E, Witten I H. Generating accurate rule sets without global optimization [C] // **The Fifteenth International Conference on Machine Learning**. Wisconsin: Morgan Kaufmann, 1998: 144-151.

[19] Gaines B R, Compton P. Induction of ripple-down rules applied to modeling large databases [J]. **Journal of Intelligent Information Systems**, 1995, **5**(3): 211-228.

[20] Witten I H, Frank E. **Data Mining: Practical Machine Learning Tools and Techniques** [M]. San Mateo: Morgan Kaufmann, 2005.

[21] Demsar J. Statistical comparisons of classifiers over multiple data sets [J]. **Journal of Machine Learning**, 2006, **7**(1): 1-30.

Fuzzy classification algorithm based on fuzzy concept similarity and fuzzy entropy measure

FENG Xing-hua[1], LIU Xiao-dong*[1], LIU Ya-qing[2]

(1. School of Control Science and Engineering, Dalian University of Technology, Dalian 116024, China;
2. School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: A method to construct a fuzzy concept similarity and fuzzy entropy measure-based classifier by using the axiomatic fuzzy set (AFS) theory is developed. A selection index based on fuzzy concept similarity and fuzzy entropy measure is proposed. Being guided by the selection index, the antecedents of the fuzzy classification rules are selected from the fuzzy concepts which are found when using the aggregation algorithm. And then, the obtained fuzzy rules are pruned by pruning algorithm, and the final classification rule group is obtained. The performance of the proposed classifier is compared with the results produced by 7 classifiers commonly encountered in the literatures when using eight datasets taken from the UCI Machine Learning Repository. It has been found that the accuracy on test data produced by the proposed classifier is higher than that produced by the other classifiers.

Key words: classification; fuzzy rules; similarity; fuzzy entropy; axiomatic fuzzy set